Character Analysis and Numerical Computations of standard M. I. Probability Distributions

Charakteranalyse und Numerische Berechnungen der standard M. I. Wahrscheinlichkeitsverteilungen

Dissertation zur Erlangung des naturwissenschaftlichen Doktorgrades der Bayerischen Julius- Maximilians- Universität Würzburg

vorgelegt von

Surath Sen
aus
Würzburg

May 18, 2013

Acknowledgements

First of all I would like to thank Prof. Dr. Manfred Dobrowolski most heartily for supervising me to the completion of my work. His priceless assistance and suggestions enabled me to make doubly sure that the software programs developed by me have running qualities more than perfect. Since the development of the software programs by means of numerical mathematical methods is the main target of my dissertation, his kind hearted full support enabled me to present my dissertation formally. Not only as a teacher, I do mightily respect him as a person. He is an extremely broad hearted person, who not only believes in fairness and justice, but also lays extreme justifiable importance to the future career life of his students.

My special thanks is addressed to Prof. Dr. Rainer Göb, who supervised the first part of my thesis involving mathematical statistics beautifully. Since the software programs developed by me give mathematical statistical models as outputs, his suggestions with regard to the presentation of statistical models were simply valuable.

My thankful gratitude is addressed to Prof. Dr. Michael Falk, whose guidance enabled me to fill up certain existent but overlooked gaps in the dissertation.

I would like to pay my hearty thanks to my father, Mr. Milan Kumar Sen, my younger brother, Sudeepto, and my uncle, Prof. Dr. Amiya Kumar Sen (*The present chairman of the department of electrical engineering, Columbia University, New York*) for giving me the basic needs as well as the basic support that played a vital and a decisive role in seeing my work to completion. My uncle, being a senior academician, has supervised at least dozens of doctoral students and is fully aware of the existing academic norms all over Europe and United States of America.

Lastly, I would also like to thank Prof. Dr. Elart von Collani for giving me certain important suggestions that contributed richly to my work. His kind assistance helped me to get my visa problems resolved. Had my visa problems been unresolved, the degree of difficulty in completing my thesis might have been beyond description.

Contents

I Elements of Stochastics 11

1 Aim of the thesis and introductory concepts 13
- 1.1 Motivation by exemplification . . . 13
- 1.2 The general idea . . . 17
- 1.3 The needed quantitative information . . . 18
- 1.4 The preliminary idea of the minimum information principle . . . 20
- 1.5 The Bernoulli space . . . 24
- 1.6 The parameter space and its elements . . . 26
- 1.7 The minimum information selection principle . . . 29
- 1.8 The formulation of the targeted aim . . . 31

2 The random structure function $\mathcal{P}$ 35
- 2.1 The random variable Y . . . 35
- 2.2 The random structure function . . . 36
- 2.3 Set partitioning of probability describing functions . . . 38
- 2.4 Efficiency of the exponential polynomial distribution . . . 39
  - 2.4.1 Exact representation of $f_Y(y)$ for the discrete case . . . 41
  - 2.4.2 Approximative representation of $f_Y(y)$ for the continuous case . . . 42
  - 2.4.3 Short summary of discrete and continuous representations . . . 45
  - 2.4.4 Continuous case as the case of approximation . . . 46
  - 2.4.5 An informatory observation . . . 48
  - 2.4.6 Conclusive points . . . 49
- 2.5 The families of probability distributions . . . 52
  - 2.5.1 The constant family: $\boldsymbol{\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \to \mathbb{P}_0}$ . . . 52
  - 2.5.2 The monotone family: $\boldsymbol{\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \to \mathbb{P}_1}$ . . . 53

2.5.3 The uni-extremal family: $\boldsymbol{\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_2}$ 55
2.5.4 A multi-extremal family: $\boldsymbol{\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_m}$ with $\boldsymbol{m > 2}$ 58

3 The two correlated selection principles 59
3.1 The stochastic entropy 59
3.2 The Shannon's measure of entropy 61
3.2.1 The Shannon's postulates 61
3.2.2 The lemma for the Shannon's theorem 62
3.2.3 The Shannon's theorem 68
3.2.4 Role of Shannon's four postulates 83
3.2.5 A property of a Shannon's postulate 84
3.2.6 Shannon's entropy as the information content 85
3.3 The principle of maximum entropy 87
3.3.1 The maximum entropy probability distribution 88
3.3.2 The Kullback-Leibler measure of deviation 94
3.3.3 A preliminary statement pertaining to the characteristic properties of λ_i, $i \in \{0, 1, 2, \ldots, m\}$ values 98
3.3.4 Characteristic properties of λ_i, $i \in \{0, 1, 2, \ldots, m\}$ values 99
3.3.5 Supporting numerical examples 106
3.3.6 The existing classical moment problems 110
3.4 The consistent density estimator 111
3.4.1 The background 111
3.4.2 The first lemma for the consistency 112
3.4.3 The second lemma for the consistency 114
3.4.4 The consistency character 115
3.5 The minimum information selection principle 120
3.5.1 Constant family: $\boldsymbol{\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_0}$ 121
3.5.2 Monotonic family: $\boldsymbol{\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_1}$ 121
3.5.3 Uni-extremal family: $\boldsymbol{\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_2}$ 122
3.5.4 The basis of the computation of λ_i, $i \in \{1, 2, \ldots, m\}$ values 123
3.6 Conclusive remarks 123
3.6.1 The monotone and uni-extremal probability distributions 123
3.6.2 A general important remark 125
3.7 The simplicity in the Hausdorff's infinite moment problem . . 126
3.7.1 The simplicity of Hausdorff's moment problem 127

4 General m. i. probability distributions 129
4.1 Introductory statements 129
4.2 The systems of simultaneous equations involving moments . . 131
4.3 Uniqueness of the solution of the equation-system for $m \in \mathbb{N}$. 134
4.4 Existence of the solution of the equation-system for $m \in \mathbb{N}$ . . 136
4.4.1 Hausdorff's finite moment problem 137
4.4.2 The necessary and sufficient condition 138
4.5 The necessary criterion for the solvability 144
4.5.1 The question addressed to the solvability 144
4.5.2 The first lemma addressed to a covariance matrix . . . 147
4.5.3 The second lemma addressed to the positive definiteness of $\mathsf{Cov}[\mathbf{X}_{(m)}]$ 149
4.5.4 The formulation of the solvability criterion 154
4.5.5 Special consideration of the cases for $m \in \{0, 1, 2\}$ . . . 156
4.6 The special Hankel matrix 158

5 Suitable simple bounds of moments 161
5.1 Basic statements for the moments of X 161
5.2 The formulation of the first target 163
5.2.1 The background 163
5.2.2 The first target 164
5.3 The formulation of the second target 167
5.3.1 The background 167
5.3.2 The second target 167
5.4 Essential role of the bounded range 168
5.4.1 The inequality defining the bounded range 169
5.4.2 The monotonic character of the bounded range 171
5.5 The targeted smallness of the bounded range 173
5.6 Existing difficulties in the discrete cases of X 175
5.6.1 The restatement of the inequality $\mu_n < \mu_{n-1}$ for a discrete X 176
5.6.2 The reestablishment of the inequality $\mu_n > \frac{\mu_{n-1}^2}{\mu_{n-2}}$ for a discrete X 177
5.6.3 General remarks 178
5.7 Bounds of μ_2 for a discrete X 179
5.7.1 The $lub(\mu_2)$ for a discrete X 179
5.7.2 The $glb(\mu_2)$ for a discrete X 180
5.8 Summary pertaining to the bounds of μ_2 183

5.9 Bounds of the first two moments of Y . . . 184
5.9.1 Upper and lower bounds of the first and the second moment of Y . . . 184
5.9.2 Upper and the lower bound of the second central moment of Y . . . 185
5.10 A preliminary note on the random variable X . . . 187

6 Standard m. i. probability distributions 189
6.1 Preliminaries . . . 190
6.1.1 The basic idea . . . 190
6.1.2 The fulfillment's plausibility check . . . 190
6.1.3 Dealing with the special cases . . . 193
6.2 M. i. monotone probability distributions . . . 194
6.2.1 Lemma for the monotonicity . . . 194
6.2.2 The restatement subjecting to the uniqueness . . . 194
6.2.3 The existence of the solution of equation-system for $m = 1$ . . . 195
6.2.4 Characteristics of the density curves . . . 197
6.2.5 Probability distributions represented by boundary points defined by $\mu_1 \in \partial\mathbf{D}^1$ . . . 197
6.3 M. i. uni-extremal probability distributions . . . 198
6.3.1 First lemma for the uni-extremity . . . 198
6.3.2 Second lemma for the uni-extremity . . . 200
6.3.3 Third lemma for the uni-extremity . . . 202
6.3.4 Fourth lemma for the uni-extremity . . . 203
6.3.5 Fifth lemma for the uni-extremity . . . 205
6.3.6 Sixth lemma for the uni-extremity . . . 206
6.3.7 Seventh lemma for the uni-extremity . . . 206
6.3.8 Eighth lemma for the uni-extremity . . . 208
6.3.9 Ninth lemma for the uni-extremity . . . 209
6.3.10 The existence of the solution of equation-system for $m = 2$ . . . 210
6.3.11 Classification of types of probability distributions . . . 230
6.3.12 Marginal variances of probability distributions . . . 231
6.3.13 Characteristic behavior of the extremal point . . . 241
6.3.14 Graphical illustrations . . . 258
6.3.15 The monotonic character as special cases . . . 261
6.3.16 Summary of probability density curves . . . 264

6.3.17 The characterizing expression 266
6.3.18 Probability distributions represented by boundary points defined by $(\mu_1, \mu_2) \in \partial\mathbf{D}^2$ 270
6.3.19 Usages of uni-extremal probability distributions 271

7 Standard distributional parameters 273
7.1 Monotone probability distribution 276
7.1.1 Discrete monotone probability distribution 276
7.1.2 Continuous monotone probability distribution 287
7.2 Uni-extremal probability distribution 293
7.2.1 Discrete uni-extremal probability distribution 293
7.2.2 Continuous uni-extremal probability distribution 297

8 Illustrations of m. i. probability distributions 307
8.1 Examples of minimum information probability distributions . 308
8.1.1 Discrete cases 310
8.1.2 Continuous cases 327
8.1.3 Conclusive remarks 349

9 Comparative studies of beta distributions 351
9.1 The beta distribution 351
9.1.1 A brief introduction 351
9.1.2 The distributional moments 352
9.1.3 The entropy 352
9.2 Generalities 355
9.3 The entropy 357
9.4 Skewness 365
9.5 Kurtosis 377
9.6 Maximum and minimum differences 391

10 Needlessness of moments more than two 417
10.0.1 A restatement regarding higher moments 418
10.0.2 Revisiting the restriction $\frac{\mu_{n-1}^2}{\mu_{n-2}} < \mu_n < \mu_{n-1}$, $n \in \mathbb{N}$ . . . 419
10.0.3 Numerical examples 421
10.0.4 Conclusions 427

II Numerical Computations 429

11 The formulation of the ultimate problem 431
11.1 Standardization of the variability 432
11.2 The transformed deterministic variable 433
11.3 The computation strategy 434
11.4 Discrete uniform probability distribution 436
11.4.1 The general case with $N > 1$ 436
11.4.2 Special case: $N = 1$ 436
11.5 Continuous uniform probability distribution 436
11.6 Discrete monotonic probability distribution 437
11.6.1 The general case with $N > 2$ 437
11.6.2 The case for constancy 438
11.6.3 The trivial case: $N = 1$ 438
11.6.4 Special case: $N = 2$ 438
11.7 Continuous monotonic probability distribution 439
11.7.1 The general case 439
11.7.2 The case for constancy 440
11.8 Usage of the standard normal density 441
11.9 Discrete uni-extremal probability distribution 446
11.9.1 The general case with $N > 3$ 446
11.9.2 The case for constancy 448
11.9.3 The monotonic case 448
11.9.4 The trivial cases: $N = 1$ and $N = 2$ 449
11.9.5 Uni-extremal probability distribution with a small variance 449
11.9.6 Special case: $N = 3$ 453
11.10Continuous uni-extremal probability distribution 456
11.10.1 The general case 456
11.10.2 The cases for symmetry and constancy 457
11.10.3 The monotonic case 458
11.10.4 Uni-extremal probability distribution with a small variance 459
11.11The overflow and underflow errors 464

12 Numerical algorithms 465
12.1 Numerical integration by Weddle's rule 466
12.1.1 The plan of action 466

12.1.2 Newton's forward interpolation formula: 470
12.1.3 Weddle's rule of numerical integration 472
12.1.4 Estimation of error in the Weddle's integral 476
12.2 Numerical solution of an equation 479
12.2.1 Iterative procedure 479
12.2.2 Condition for the convergence of the iterative procedure 481
12.2.3 Rate of convergence of the iterative procedure 481
12.2.4 Newton Raphson procedure 482
12.2.5 Condition for the convergence of the Newton Raphson procedure 483
12.2.6 Rate of convergence of the Newton Raphson procedure 484
12.2.7 Complete solution of the equation $f(x) = 0$ 484
12.3 Numerical solution of a system of two equations 487
12.3.1 Newton Raphson procedure 487
12.3.2 The lemma for the convergence of the Newton Raphson procedure 490
12.3.3 Condition for the convergence of the Newton Raphson procedure 492
12.3.4 Role of the Newton Raphson's convergence criterion in special cases 500
12.3.5 Iterative procedure for two special equation-systems . . 507
12.3.6 Conditions for the convergence in cases of the two special systems 510
12.3.7 Newton Raphson procedure for the two special equation-systems 515
12.4 Discrete monotonic probability distribution 519
12.4.1 Algorithmic steps 519
12.4.2 The subfamily 521
12.5 Continuous monotonic probability distribution 522
12.5.1 Algorithmic steps 522
12.5.2 The subfamily 524
12.6 Discrete uni-extremal probability distribution 525
12.6.1 Algorithmic steps 525
12.6.2 The subroutine for non special uni-extremal cases . . . 531
12.7 Continuous uni-extremal probability distribution 534
12.7.1 Algorithmic steps 534
12.7.2 The input- and the solution space 538
12.7.3 Characteristics of the outputs with respect to the inputs 539

12.7.4 An useful transformation $Z = 1 - X$ 542
12.7.5 The data structure 546
12.7.6 Database access procedures for start vectors 547
12.7.7 Algorithm for subsequent processing of access vectors . 560
12.7.8 Modified iterative procedures 569
12.7.9 The subroutine for non special uni-extremal cases . . . 572
12.8 Limitations 576
12.8.1 Probable limitations of my software programs in monotone cases 576
12.8.2 Probable limitations of my software program in discrete uni-extremal cases 576
12.8.3 Limitations of my software program in continuous uni-extremal cases 576
12.9 Operating instructions 578

A The maximization of the stochastic entropy **581**
A.1 The expression of the maximum entropy probability distribution 581

B The role of the Hankel matrix **583**
B.1 The important lemma 583
B.2 Uniqueness of the solution of the equation-system for $m \in \mathbb{N}_0$. 585

C Miscellaneous **591**
C.1 Inequalities 591

Part I

Elements of Stochastics

Chapter 1

Aim of the thesis and introductory concepts

1.1 Motivation by exemplification

This thesis is principally devoted to the construction of **appropriate approximating** probability distributions, which are made to serve certain essential needs in accordance with the situations. These aforesaid probability distributions are constructed by the **minimum information principle**. This principle shall be discussed elaborately in due course. The importance of this work can be well motivated by picturing the following particular practical problem: The natural wind exerts a force on the rotor blades of a wind turbine. This force may be termed as the wind-load on the rotor blades. Depending on how large the wind-load could be, we need to determine the maximum load on every rotor blade termed as maximum rotor-load. This maximum rotor-load is **exclusively** caused by the wind-load and no other external factors are taken into consideration. In our task, this maximum rotor-load is defined by the continuous random variable Y_w. The probability distribution of Y_w, which plays a deciding role here and which is determined by the minimum information principle, shall also be discussed here.

For the purpose of designing or constructing the rotor blades, the knowledge about this maximum rotor-load is of utmost importance.

The problem is, the maximum rotor-load is generally unknown and therefore needs to be investigated by statistical methods. Of

course, this investigation is subjected to certain specified risks.

So, we are going to go ahead with the discussion of this particular problem. For the sake of simplicity, it is assumed that the **weather conditions are constant**, otherwise we shall need go for unnecessary complications.

The natural wind causing the wind-load exerted on the rotor blades has certain speeds. This wind speed has a certain variability. Realistically speaking, this variability has to be a bounded range. In the language of Statistics, this wind speed can be denoted by the continuous random variable W. Referring to the source [10], the behavior of W is described by

- $v_w = E[W]$ = Average wind speed
- $\sigma_w = \sqrt{E[(W - v_w)^2]}$ = Turbulence (standard deviation)

With subject to the informative fact that σ_w is not unbelievably large, we can justify the assumption that the probability distribution of W is given by a bell- shaped [1] probability density curve. *The question of the relationship between the smallness of σ_w and the bell- shapeliness of the probability distribution under consideration shall be discussed in due course.*

At this point, let us take exactly one rotor blade of the wind turbine at random for our consideration.

Again, the wind-load on this rotor blade (i.e. in other words, the influence of wind on the rotor blade) changes with the position of this rotor blade. The aspect of our interest shall be the **maximum wind-load**, i.e. principally the **maximum rotor load**. Thus, there exists position of this rotor blade, where the rotor load is maximum. Experiments have been carried out and it has been found that this maximum rotor load shows an inherent variability and this particular statement is referred to the source [42].

It is intuitively clear that this maximum rotor load is directly proportional to the wind speed. It can be shown that this **maximum rotor load** described by Y_w has the same type of probability distribution as the same of W (*the formal argument of this very statement is beyond the scope of this work*). In other words, the probability distribution of Y_w is of uni-modal type, i.e. the probability density curve of Y_w is bell-shaped.

[1] in the language of Statistics, a bell- shaped probability distribution is called uni-modal probability distribution

Again, from the realistic point of view, the probability distribution of Y_w has to have a bounded support[2]. Consequently, we shall proceed to construct the probability distribution of Y_w by the minimum information principle, after having estimated the following parameters:

- $a = \min[Y_w]$. Usually, we take $a = 0$.
- $b = \max[Y_w]$
- $\mu^{(1)}_{Y_w} = E[Y_w]$
- $\sigma^2_{Y_w} = Var[Y_w]$

By this principle, the probability density function of Y_w denoted by $f_{Y_w}(y)$ is given as

$$f_{Y_w}(y) = e^{\lambda_0+\lambda_1 y+\lambda_2 y^2}, \;\; a \leq y \leq b$$

where the values of λ_1 and λ_2 are uniquely given by $\mu^{(1)}_{Y_w}$ and $\sigma^2_{Y_w}$.

As a matter of fact, the parameters $\mu^{(1)}_{Y_w}$ and $\sigma^2_{Y_w}$ are not estimated only once in practice, but certain number of times for the purpose of achieving a reasonably good picture of the whole thing. But, for the sake of simplicity, we shall show by a simple example, how the density function $f_{Y_w}(y)$ can be found out and how the maximum rotor-load can be found out subsequently.

Let the estimated values of a and b be 0 and 12000 units respectively. This necessarily means that any possible value of Y_w exceeding 12000 or falling below 0 is completely ruled out. Moreover, let the estimated values of $\mu^{(1)}_{Y_w}$ and $\sigma^2_{Y_w}$ be 7500 and 170000 (i.e $\sigma_{Y_w} = 412.311$ units) respectively. With subject to these data, $f_{Y_w}(y)$ is given as

$$f_{Y_w}(y) = e^{-172.38189186180912+0.04411764705882353y-2.941176470588235510^{-6}y^2}, \\ 0 \leq y \leq 12000$$

Clearly, the density function $f_{Y_w}(y)$ is bell-shaped.

[2]According to the IEC-Standard, the probability distribution of Y_w is taken to be a Weibull distribution with unbounded support. In our discussions, we shall use the estimated parameters to construct the probability distributions with bounded support

Consequently, if the involved risk in determining the maximum rotor-load is taken to be 2.5 %, then we can easily achieve the following

$$\int_{6575.85}^{8424.15} f_{Y_w}(y)dy = 0.975 = 1 - 0.025$$

which says that the desired maximum rotor-load lies within the interval $[6575.85, 8424.15]$ (i.e. lies in between 6575 units and 8424.15 units), with subject to the risk of 2.5 %.

Notably, this closed interval, namely $[6575.85, 8424.15]$, which contains the desired maximum rotor-load with subject to the risk of 2.5 %, has been computed in a way that the **length of the interval must be minimum**. In this case, this length is 1848.3 units.

This computed closed and bounded interval, in the language of stochastics (referred to page 233 of [39] or to the detailed explanations given in [38]), is called the **optimal prediction interval** with a **reliability level** [3] of 0.975. This prediction interval (or prediction) (with reference to the page 233 of [39]) is symbolized as $A_Y^{(0.975)}(\{(7500, 56420000)\}) = [6575.85, 8424.15]$. The figures $\mu_{Y_w}^{(1)} = 7500$ and $\mu_{Y_w}^{(2)} = \left(\mu_{Y_w}^{(1)}\right)^2 + \sigma_{Y_w}^2 = 56420000$ are the estimated values of the first two moments of Y respectively. This is precisely the way, how such a prediction interval is computed with respect to a given specified risk. However, if the length of this interval is not minimum, then it is a prediction interval with subject to the given reliability specification though, but **not optimal**. This kind of determination of this interval demands that the probability distribution of Y_w needs to be a bell-shaped one.

In case we intend to reduce the risk to 1 %, then the upper limit of the maximum rotor-load is found to be increased to 8562.04 units and the lower limit of the maximum rotor-load to be reduced to 6437.96 units. In that case, the optimal prediction interval with a reliability level of 0.99 becomes $A_Y^{(0.99)}(\{(7500, 56420000)\}) = [6437.96, 8562.04]$. This basically says that if the involved risk is reduced completely to 0 %, then the maximum rotor-load is nothing different from 12000 units and thereby reducing the risk to 0 % makes such stochastic procedures completely redundant.

[3] the chosen reliability level of a stochastic procedure ranges from 0 to 1 (referred to the page 19 of [53] as well as to the current [38])

Thus, taking reasonable risks makes the stochastic procedures worthy.

This way of solving this pictured practical problem should set us thinking that we need to construct appropriate **situation oriented** probability distributions for the purpose of performing certain important needful tasks. In other words, this should motivate us to go deeper into the search of finding the appropriate approximating probability distributions.

The approximating probability distributions are those stochastic models, which are principally used to develop **stochastic procedures**, for eg. **prediction procedures**.

1.2 The general idea

Almost every probability distribution necessitates the knowledge of certain relevant parameters. These parameters are the usually termed as distributional parameters. Therefore, a probability distribution of a specified type can be determined by suitably chosen parameters. The number of parameters as well as the parameters themselves determine the complexity of the probability distribution and the type of the probability distribution is termed accordingly. The type of the probability distribution is ascribed to the uniform, monotonic, uni-extremal and multi-extremal nature of probability distributions. We shall categorize these types in accordance with the number of extremes of the probability mass function or the probability density function, as the case may be.

The random variable, say Y, whose probability distribution is of interest, basically describes a random size of a particular aspect of interest. If this aspect of interest has to tally with the real world, the random size cannot be of infinite size. This must explain, why the support of the probability distribution of Y (i.e. the range of variability of Y) must be either finite, if Y is of discrete type or a closed and bounded interval, if Y is of continuous type.

Quite frequently, the cases do arise, when the exact function describing the probability distribution of a specified type (i.e. an user-given type) is an open question. In that case, the probability distribution fitting to the particular situation under consideration needs to be constructed by certain rules and principles. This construction technique needs to be specifically defined by

mathematical means and implemented for practical needs. The formulation and the implementation of this construction technique is precisely the aim of this dissertation. However, the implementation is restricted to the commonly known cases, namely uniform, monotonic and uni-extremal cases, which shall be elaborated in due course.

This construction technique necessitates an amount of information about the probability distribution of the specified type. So, we now proceed to elaborate, what we do exactly mean by the amount of information necessary to construct the probability distribution of the (user) specified type.

1.3 The needed quantitative information

In every given situation, the user has to specify the type of the probability distribution he needs for his requirements. For eg., if he needs to play or conduct a game of dice or any game of pure chances, he needs a *constant discrete probability distribution*; if he needs to draw conclusions out of a result of a random experiment by performing a specified finite number of Bernoulli trials (i.e. each trial has a fixed probability of success), he needs a *uni-modal discrete probability distribution*, etc. Coincidentally, this particular uni-modal probability distribution is known to be the binomial distribution.

However, in each given situation, the user may or may not be provided with a probability distribution that is conventionally known. So, in case the probability distribution is not available at hand, it needs to be constructed by means of the available quantitative knowledge describing the situation.

This very concept of **available quantitative knowledge** is interpreted mathematically as **the quantitative information needed** to determine the desired probability distribution of Y uniquely. As a matter of fact, even if the function specifying the probability distribution of Y in the given situation under consideration is not known, the **distributional moments** of Y are always **empirically estimable** and therefore the utilization of these **moments as the provided quantitative information** for the probability distribution is imaginable. For this, we need a good amount of work to perform.

Moreover, it is well understood that the point estimation procedures or interval estimation procedures for estimating the parameters in form of moments

are **relatively simpler** in comparison to the estimation of parameters in other forms (i.e. in forms different from the moment-form of parameters).

The theoretical probability distributions are conventionally known to be **based on certain mathematical rules governing the laws of probability** only. Unlike the conventional theoretical probability distributions, for our practical purposes,

- the selection of the *situation oriented need based* probability distribution is **based on the minimum information principle**.
- the construction (development) of the *selected situation oriented need based* probability distribution is strictly **based on empirical experiences**. One of the reasons for this is, that the estimation procedures for estimating the moments are principally based on the **empirical experimental results**.

Let us denote the **existing** but generally **unknown situation oriented need based** probability distribution of Y having the support $\mathcal{X}_Y$ by

$$f_Y(y),\ y \in \mathcal{X}_Y \tag{1.1}$$

The function $f_Y(y)$ is the **probability mass function** or the **probability density function**, according as Y is discrete or continuous.

Purely theoretically speaking, the **complexity** of a probability distribution of a **specified type** is not necessarily fixed and is **ascribed** to the **amount of quantitative information**[4]. So, from the theoretical point of view, if the amount of information necessary to construct a probability distribution of the desired (specified) type has to be reduced to minimum, then the complexity of the probability distribution of specified type should also turn out to be minimum. This minimum amount of information can be defined by the **minimum number of necessary distributional moments** (referred to the page 167 of [54]) for the construction of the probability distribution of a specified type, in addition to the preassigned support of the probability distribution.

It has to be carefully noted that the utilization of the amount of information more than the minimum necessary information means an increase in accuracy

[4]From now on, for the sake of simplicity, we shall term the **quantitative information** simply as **information**

from the **theoretical point of view** only. But, from the **practical point of view**, it is rather harmful, simply because only the estimated values (not the exact theoretical actual values) of the moments of Y are taken into consideration. Moreover, this too would mean a significant increase in amount of programming work.

Keeping this in mind, we shall engage ourselves in determining (constructing) the probability distribution of the specified type with subject to the **minimum information** needed for the construction. The construction of the required probability distribution of specified type is the basic aim of this thesis.

However, every required probability distribution does not have the same degree of complexity. So, this aim necessitates an intensive study of the degree of complexity of each type of probability distribution that can be specified by the user, especially those types that are of frequent use.

1.4 The preliminary idea of the minimum information principle

Before we go ahead, we need to briefly state that the support of the probability distribution of Y and the moments of Y are denoted by $\mathcal{X}_Y$ and $\mu_Y^{(i)}$, $i \in \mathbb{N}$ respectively.

The constructed probability distribution of the specified type with the help of the minimum possible information is called the minimum information probability distribution and the principle of construction involved in this regard, is termed as the **minimum information principle**. This is to say that under the consideration of the family of all the probability distributions with bounded support, the minimum information principle says that in a given situation if a probability distribution of a particular type is needed, then, only that particular member of the family has to be identified which needs the minimum quantitative information.

This idea of minimum information for the selection of a specified type of probability distribution of Y is sketched as follows:

If the specified type of probability distribution is

1. uniform (constant), then the knowledge of $\mathcal{X}_Y$ is needed as the information at the **very least**.

2. monotone, then the knowledge of $\mathcal{X}_Y$ and that of the first moment, namely $\mu_Y^{(1)}$, is needed as the information at the **very least**.

3. uni-extremal, then the knowledge of $\mathcal{X}_Y$ and that of the first two moments, namely $\mu_Y^{(1)}$ and $\mu_Y^{(2)}$, is needed as the information at the **very least**.

4. $(m-1)$- extremal ($m \in \mathbb{N}$, $m > 2$), i.e. multi-extremal with $m-1$ extremes, then the knowledge of $\mathcal{X}_Y$ and that of the first m moments, namely $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}$, is need as the information at the **very least**.

Basically, the minimum information principle says that exclusively the **key moments** [5] are utilized for the construction of the probability distribution of the user-specified type. The concept *key moments* means the moments of dire necessity. This is to say that for a probability distribution with $m-1$ extremes ($m \geq 1$), the key moments are $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}$, whereas the other moments $\mu_Y^{(m+1)}, \mu_Y^{(m+2)}, \ldots$ are the non-key moments.

Now, the next question arises, how to select the probability distribution of Y of a specified type with subject to the given $\mathcal{X}_Y$ and the moments of Y mathematically? The answer to this question is, the selection of that particular probability distribution of Y with subject to the given $\mathcal{X}_Y$ and, if necessary, the moments $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}$, $m \geq 1$, so that the probability distribution has the maximum entropy. In other words, this selection is based on the **maximum entropy principle**. Referring to the page 1 of [25], this principle states that, with subject to a given set of probability distributions of Y, each element of **which is consistent with the specified information about the probability distribution** (for eg. the information stating predeterminately the support $\mathcal{X}_Y$ and probably the moments $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}$, $m \geq 1$ of the probability distribution of Y of necessity), the user selects that particular probability distribution of Y from the set that has the **maximum entropy**.

[5] The German translation of **key moments** is Schlüsselmomente, the concept of minimum information in German says "nur die **notwendigste** Anzahl der Momente sind erforderlich"

Notably, the case of $m = 0$ is included here. The means: The information about the probability distribution includes the support $\mathcal{X}_Y$ only, i.e. not the moments of Y. In that case, this information says that the probability distribution of Y is exclusively uniform.

As a matter of fact, referring to the system of equations (4.2), it has been duly shown in the subsection 3.3.1 that, with subject to the given $\mathcal{X}_Y$ and $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}$, $m \geq 1$ in form of the information about the probability distribution of Y,

- for a discrete Y, the probability mass function of Y, namely

  $$f_Y(y_j) = P_Y(\{y_j\}) = e^{\lambda_0 + \lambda_1 y_j + \lambda_2 y_j^2 + \ldots + \lambda_m y_j^m},\ j \in \{1, 2, \ldots, N\} \quad (1.2)$$

  has the maximum entropy, where $\mathcal{X}_Y = \{y_1, y_2, \ldots, y_N\}$ and $a = y_1 < y_2 < \ldots < y_N = b$

- for a continuous Y, the probability density function of Y, namely

  $$f_Y(y) = e^{\lambda_0 + \lambda_1 y + \lambda_2 y^2 + \ldots + \lambda_m y^m},\ y \in \mathcal{X}_Y \quad (1.3)$$

  has the maximum entropy, where $\mathcal{X}_Y = [a, b]$ and this means that the resulting value of Y is unlikely to fall below a and to exceed b.

and in both discrete and continuous cases, the aforesaid representations of both $P_Y(\{y_j\})$ and $f_Y(y)$ can be **determined uniquely** (referred to the section 4.3) **under certain existence conditions** (referred to the section 4.4). These existence conditions involve the moments in both the discrete and continuous cases of Y, but involve the support $\mathcal{X}_Y$ in discrete cases of Y only.

Because of this very **uniqueness** of these aforesaid representations, namely (1.2) for a discrete Y and (1.3) for a continuous Y, we are now in a position to state that the moments of Y, namely $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}$, can be treated as the **parameters of the probability distribution** of Y.

Therefore, we need to define a set containing such parameters in a way that each of its elements of it determines a probability distribution of Y uniquely. This set may be termed as the **parameter space**. In the language of stochastic science, this parameter space is called the **ignorance space**, symbolized by $\mathcal{D}_Y$. This is basically to say that, each element of the set $\mathcal{D}_Y$ together with

the preassigned support $\mathcal{X}_Y$ of the probability distribution must determine a particular member of the family of probability distributions of Y.

At this point, we need to introduce the mathematical structure **Bernoulli space** $\mathbb{B}_{Y,d_Y}$ describing a real natural phenomenon of interest (can be referred to the page 155 of [54] as well as to the e-learning programme [38]). The inherent randomness present in any real natural phenomenon is described by the random structure of the random variable Y, the key parameters of the probability distribution of Y being given by d_Y. The mathematical structural representation of d_Y shall be described shortly.

But, for this, we need to state preliminarily in advance, that, since we have already stated that the key moments of Y can be treated as the parameters of the probability distribution of Y, d_Y is a representation of the key moments of Y. So, we denote the constructed probability distribution (i.e. the approximating probability distribution) of Y with the help of the key parameters (i.e. key moments) duly defined by d_Y as

$$f_{Y|\{d_Y\}}(y),\ \ y \in \mathcal{X}_Y \tag{1.4}$$

the support of the probability distribution of Y being denoted by $\mathcal{X}_Y$.

In a simple language, the probability distribution of Y given by (1.4) is the **approximating** probability distribution of the actual, existing but unknown probability distribution of Y given by (1.1).

The reason for the usage of the set notation for denoting the probability distribution of Y as dependent on d_Y by putting $\{d_Y\}$ instead of putting d_Y in the representation (1.4) shall be briefed shortly.

The probability distribution of Y given by (1.4) stands for both the **discrete** and **continuous** cases of Y. However, if we intend to be specific about the discrete case of Y, just for a better degree of clarity, we shall denote the probability mass function of Y as

$$P_{Y|\{d_Y\}}(\{y\}),\ \ y \in \mathcal{X}_Y \tag{1.5}$$

With this, we proceed to introduce the Bernoulli space.

1.5 The Bernoulli space

The stochastic model Bernoulli space, symbolized by $\mathbb{B}_{Y,d_Y}$, (referred to the page 155 of [54]) is a **three tuple** [6] mathematical algebraic structure. This $\mathbb{B}_{Y,d_Y}$ pictures the **unknownness** of d_Y (i.e. the **ignorance** about d_Y) as well as the **randomness** of the random variable Y, this randomness being controlled largely by d_Y.

This randomness of Y is described by the probability distribution of Y and is therefore of basic interest. So, our objective shall be to construct this ***situation oriented need based*** probability distribution of Y by means of our available information.

The first, the second and the third element of $\mathbb{B}_{Y,d_Y}$ are termed as the **ignorance space** (symbolized by $\mathcal{D}_Y$), the **variability function** (symbolized by $\mathcal{X}_Y\left(\mathcal{D}_Y^{(0)}\right)$) and the **random structure function** (symbolized by $\mathcal{P}\left(\mathcal{D}_Y^{(0)}\right)$) respectively, such that

- $\mathcal{D}_Y^{(0)}$ denotes any subset of $\mathcal{D}_Y$ i.e. $\mathcal{D}_Y^{(0)} \subseteq \mathcal{D}_Y$.
- both the variability function and the random structure function are the functions of the subsets of $\mathcal{D}_Y$, i.e. of $\mathcal{D}_Y^{(0)}$

The Bernoulli space is thereby structurally defined by

$$\mathbb{B}_{Y,d_Y} = \left(\mathcal{D}_Y, \mathcal{X}_Y\left(\mathcal{D}_Y^{(0)}\right), \mathcal{P}\left(\mathcal{D}_Y^{(0)}\right)\right) \tag{1.6}$$

where d_Y denotes any element of $\mathcal{D}_Y$, viz. $d_Y \in \mathcal{D}_Y$, or in other words, $\{d_Y\} \subseteq \mathcal{D}_Y$.

Here, the **domain of definition** of both the variability function $\mathcal{X}_Y\left(\mathcal{D}_Y^{(0)}\right)$ and the random structure function $\mathcal{P}\left(\mathcal{D}_Y^{(0)}\right)$ is a **suitably chosen** system of subsets of $\mathcal{D}_Y$ denoted by $\mathcal{T}_{D_Y}(\mathcal{D}_Y)$. The choice of $\mathcal{T}_{D_Y}(\mathcal{D}_Y)$ depends **exclusively** on the situation under consideration.

Now, let us define the three elements of the Bernoulli Space (with reference to the current magister e-learning programme [38]) briefly as follows:

[6] A tuple is a mathematical structure, which is a **ordered** list of certain elements.

Definition 1.5.1 (The ignorance space (or the parameter space). *The* ***ignorance space*** $\mathcal{D}_Y$ *consists of all the potential values of* d_Y*, which* ***cannot*** *be excluded at a given point of time.*

Definition 1.5.2 (The variability function). *The* ***variability function*** $\mathcal{X}_Y\left(\mathcal{D}_Y^{(0)}\right)$ *has an* ***argument*** $\mathcal{D}_Y^{(0)} \in \mathcal{T}_{D_Y}(\mathcal{D}_Y)$ *(for a particular* $\mathcal{D}_Y^{(0)} \subseteq \mathcal{D}_Y$*) and the* ***codomain*** *as a* ***compact set*** *of all the possible empirical values of* Y*, is described by*

$$\mathcal{X}_Y : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \to \left\{ y \left| y \in \left[\min_{d_Y \in \mathcal{D}_Y} \mathcal{X}_Y\left(\mathcal{D}_Y\right), \max_{d_Y \in \mathcal{D}_Y} \mathcal{X}_Y\left(\mathcal{D}_Y\right) \right], y \in \mathcal{X}_Y(\mathcal{D}_Y) \right. \right\} \tag{1.7}$$

and in fact

$$\mathcal{X}_Y\left(\mathcal{D}_Y^{(0)}\right) = \bigcup_{d_Y \in \mathcal{D}_Y^{(0)}} \mathcal{X}_Y(\{d_Y\}) \tag{1.8}$$

such that $\mathcal{X}_Y(\{d_Y\})$ *is the* ***compact support*** *of the probability distribution of* Y *(this compact support is a finite set in case of a discrete* Y *or a compact interval in case of a continuous* Y *).*

In particular, the ***support*** *of the probability distribution of* Y *defined by an element* d_Y *(*$d_Y \in \mathcal{D}_Y$*) denoted by* $\mathcal{X}_Y(\{d_Y\})$ *is thereby given by*

$$\mathcal{X}_Y(\{d_Y\}) = \{ y \,| y \in [\min \mathcal{X}_Y(\{d_Y\}), \max \mathcal{X}_Y(\{d_Y\})] \ , y \in \mathcal{X}_Y(\{d_Y\}) \} \tag{1.9}$$

(i.e $\mathcal{X}_Y(\{d_Y\})$ *= the* ***range of variability*** *of* Y *determined by* d_Y*)*

Definition 1.5.3 (The random structure function). *The random structure function* $\mathcal{P}\left(\mathcal{D}_Y^{(0)}\right)$ *has an* ***argument*** $\mathcal{D}_Y^{(0)} \in \mathcal{T}_{D_Y}(\mathcal{D}_Y)$ *(for a particular* $\mathcal{D}_Y^{(0)} \subseteq \mathcal{D}_Y$ *) and the* ***codomain*** *as a* ***family*** $\mathbb{P}$ *of all the possible* ***exponential polynomial probability distributions*** *of* Y *that may be brought under consideration, is described by*

$$\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \to \mathbb{P} \tag{1.10}$$

and in fact

$$\mathcal{P}\left(\mathcal{D}_Y^{(0)}\right) = \frac{\sum_{d_Y \in \mathcal{D}_Y^{(0)}} f_{Y|\{d_Y\}}(y)}{\left|\mathcal{D}_Y^{(0)}\right|}, \ y \in \mathcal{X}_Y\left(\mathcal{D}_Y^{(0)}\right) \tag{1.11}$$

such that $\left|\mathcal{D}_Y^{(0)}\right|$ *is the cardinality of the set* $\mathcal{D}_Y^{(0)}$.

In particular, the ***probability distribution*** *of* Y *defined by an element* d_Y *(*$d_Y \in \mathcal{D}_Y$*) (i.e the* ***probability mass function*** *or the* ***probability density function*** *of* Y*, according as* Y *is discrete or continuous) denoted by* $f_{Y|\{d_Y\}}(y)$ *is thereby defined by*

$$\mathcal{P}(\{d_Y\}) = f_{Y|\{d_Y\}}(y), \;\; y \in \mathcal{X}_Y(\{d_Y\}) \tag{1.12}$$

having the ***support*** $\mathcal{X}_Y(\{d_Y\})$.

As far as the **relevancy of this dissertation** is concerned, we shall restrict our discussions to **singleton subsets** of $\mathcal{T}_{D_Y}(\mathcal{D}_Y)$ **only**. In other words, we shall restrict our discussions to $\mathcal{D}_Y^{(0)} = \{d_Y\}$ only. The very fact that the subset $\mathcal{D}_Y^{(0)}$ of the ignorance space $\mathcal{D}_Y$ is an argument of both the variability function and the random structure function therefore clarifies, why we use the notation $\{d_Y\}$ instead of simple d_Y.

As the next step, we shall proceed to discuss the parameter space (or equivalently the ignorance space) elaborately.

1.6 The parameter space and its elements

As a matter of fact, the parameters must be deterministically specified to determine a probability distribution of a specified type **uniquely**. In other words, the word **deterministic nature** of a parameter means, the parametric value under consideration is either specifically known or can be suitably estimated by conventional estimation procedures. Therefore, since the parameters are needed to be of deterministic nature, the finite collection of all the necessary parameters to construct a probability distribution can be termed as the deterministic variable.

Now, let us take an example of the binomial distribution followed by the discrete random variable Y, where the distributional parameters are either known in advance or estimated to be N and p. Here, N and p determine the binomial distribution of Y uniquely. Moreover, the first two moments of the binomial distribution, namely Np and $Np(1-p)+(Np)^2$ determine the binomial distribution of Y uniquely as well. It is therefore absolutely clear that any empirically estimated values of $\mu_Y^{(1)}$ and $\mu_Y^{(2)}$ determine the binomial

distribution uniquely and therefore the deterministic variable specifying the probability distribution (namely the binomial distribution) of Y, which can be symbolized as d_Y, can be presented as $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$.

However, in case of a binomial distribution, which is uni- modal by nature, it remains to be unforgettably stated that this distribution is uniquely determinable with the help of the support $\mathcal{X}_Y = \{0, 1, \ldots, N\}$ and $\mu_Y^{(1)}$ only, i.e. even by ignoring the $\mu_Y^{(2)}$. This is a pure and simple coincidence. It must be clearly stated that, **in general**, the construction of a uni-modal probability distribution is **not possible** with the help of $\mathcal{X}_Y$ and $\mu_Y^{(1)}$ only, but an additional need of $\mu_Y^{(2)}$ is absolutely necessary. This is precisely, what the minimum information principle says.

Moreover, a very important careful observation has to be made here (i.e. in this case of binomial distribution). It can be easily seen that $\mathcal{X}_Y$ (being dependent on N) is dependent on d_Y and in fact, this dependency can be well described by the following inequalities:

- $0 \leq \mu_Y^{(1)} \leq N$
- $\left(\mu_Y^{(1)}\right)^2 \leq \mu_Y^{(2)} \leq N\mu_Y^{(1)}$

In general, it shall be shown in due course, (referred to (5.34) and (5.35)) that

- $a \leq \mu_Y^{(1)} \leq b$
- $\left(\mu_Y^{(1)}\right)^2 \leq \mu_Y^{(2)} \leq (a+b)\mu_Y^{(1)} - ab$
- every moment $\mu_Y^{(i)}$, $i \in \mathbb{N}$ is restricted by the values of the moments $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(i-1)}$ (of lower order) as well as by the values of a and b

and thus the general expression of d_Y, which defines the probability distribution of Y uniquely, is given by

$$d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}) \tag{1.13}$$

This aforesaid dependency enables to express the support $\mathcal{X}_Y$ as a function of d_Y, namely $\mathcal{X}_Y = \mathcal{X}_Y(\{d_Y\})$ (referred to the established definition (1.9)).

Now, by having a look at the expressions (1.2) and (1.3), we can well see that $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the m parameters of a probability distribution. In that case, the quantity λ, being defined by

$$\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m),$$

determines the probability distribution of Y in each of the cases (1.2) and (1.3) uniquely, provided the support $\mathcal{X}_Y$ is known too.

Importantly, in both the cases of (1.2) and (1.3), λ_0 is uniquely determinable by $\lambda_1, \lambda_2, \ldots, \lambda_m$

This enables us to conclude that with subject to a given $\mathcal{X}_Y$, either d_Y or λ determine the probability distribution of Y uniquely. In other words, the representations given by d_Y and λ are **basically equivalent**.

Therefore, our basic task of this dissertation is formulated in the following way:

- An element of the **ignorance space** (denoted by d_Y) is empirically known and the **variability function** $\mathcal{X}_Y = \mathcal{X}_Y(\{d_Y\})$ is chosen accordingly and appropriately as per (1.9).
- Immediately after this, by solving a system of simultaneous nonlinear equations for λ either analytically or numerically, thereby giving the desired probability distribution of Y, namely the **random structure function** $f_{Y|\{d_Y\}}$, $y \in \mathcal{X}_Y(\{d_Y\})$ as per (1.12).

  The **numerical computation of the desired random structure functions** by **numerical mathematical methods** had been simply a **backbreaking amount of programming work** for me.

So, as it is clear that λ (i.e. each λ_i, $i \in \{1, 2, \ldots, m\}$) is dependent on d_Y, we can express our λ in the following way:

$$\lambda(d_Y) = (\lambda_1(d_Y), \lambda_2(d_Y), \ldots, \lambda_m(d_Y)) \tag{1.14}$$

Therefore, since every empirical value of d_Y is decisively important, we need to find out the complete set containing the different possible empirical values of d_Y, simply for the purpose of knowing, which values of d_Y are at all expected or imaginable. This set is termed as **parameter space** (or **ignorance space**), symbolized by $\mathcal{D}_Y$.

In other words, the variable, whose variability range is $\mathcal{D}_Y$, is denoted by d_Y. So, we term d_Y as the **deterministic variable**, each empirical value of which determines a particular probability distribution of Y with subject to the additionally empirically known $\mathcal{X}_Y(\{d_Y\})$. Of course, d_Y can be equivalently expressed in other forms as well.

In a plain and simple language, since the representation (1.13) determines the representation (1.14) uniquely for the purpose of determining the probability mass function of Y given by (1.2) (or the probability density function of Y given by (1.3)), the representation (1.14), namely λ, is termed as the **distribution related representation of the deterministic variable**, whereas the equivalent representation (1.13), namely d_Y, is termed as the **moments related representation of the deterministic variable**.

Thus, if we intend to know the set of all possible probability distributions of Y, we need to know the $\mathcal{D}_Y$. Of course, if we find that some of the probability distributions of Y defined by $d_Y \in \mathcal{D}_Y$ are absurd or useless, then we can exclude such elements of $\mathcal{D}_Y$.

Before ending this section, we would like to state another important thing **unforgettably**. We have already mentioned that every moment $\mu_Y^{(i)}$ of Y with $i \in \{1, 2, \ldots, m\}$ of Y is restricted by upper and lower bounds and therefore cannot assume any arbitrary real values. This means that the parameter space $\mathcal{D}_Y$ is different from $\mathbb{R}^m$, each element of it, being denoted by d_Y, is described by the representation (1.13).

With this, we proceed to introduce the definition of the minimum information principle formally, but briefly.

1.7 The minimum information selection principle

In this section we shall define the minimum information selection principle (referred to the pages from 167 to 173 of [54]). The elaboration of the **clearer meaning of the concept of minimum information** shall be carried out in the section 3.5.

The selection of an appropriate probability distribution is to be performed particularly on the basis of empirical experiences. As we have already dis-

cussed, the type of the selected probability distribution is ascribed to the number of extremes ($= m - 1,\ m \geq 1$) of the probability distribution of Y (i.e. either of the probability mass function of a discrete Y or of the probability density function of a continuous Y).

The minimum information selection principle for selecting a probability distribution of Y therefore says the following:

1. The family of **constant probability distributions** is denoted by $\mathbb{P}_0$. The random structure function $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_0$ having the codomain $\mathbb{P}_0$ is nothing but a plain and simple constant function.

   That is, the value of the function $f_{Y|\{d_Y\}}(y)$ is constant throughout the range $\mathcal{X}_Y(\{d_Y\})$, where $d_Y = ()$ does not contain any moment of Y.

   The principle says that, for the purpose of constructing a **constant** probability distribution of Y, the exact available information must read: the knowledge of $\mathcal{X}_Y(\{d_Y\})$ only and not of any moments of Y.

2. The family of **monotone probability distributions** is denoted by $\mathbb{P}_1$. The random structure function $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_1$ having the codomain $\mathbb{P}_1$ is given by

   $$\mathcal{P}(\{d_Y\}) = f_{Y|\{d_Y\}}(y) = e^{\lambda_0 + \lambda_1 y},\ y \in \mathcal{X}_Y(\{d_Y\}) \tag{1.15}$$

   such that $d_Y = (\mu_Y^{(1)})$.

   The principle says that, for the purpose of constructing a **monotone** probability distribution of Y, the exact available information must read: the knowledge of $\mathcal{X}_Y(\{d_Y\})$ and of $\mu_Y^{(1)}$.

3. The family of **uni-extremal probability distributions** is denoted by $\mathbb{P}_2$. The random structure function $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_2$ having the codomain $\mathbb{P}_2$ is given by

   $$\mathcal{P}(\{d_Y\}) = f_{Y|\{d_Y\}}(y) = e^{\lambda_0 + \lambda_1 y + \lambda_2 y^2},\ y \in \mathcal{X}_Y(\{d_Y\}) \tag{1.16}$$

   such that $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$.

   The principle says that, for the purpose of constructing a **uni-extremal** probability distribution of Y, the exact available information must read: the knowledge of $\mathcal{X}_Y(\{d_Y\})$ and of $\mu_Y^{(1)}$ and $\mu_Y^{(2)}$.

4. The family of **multi-extremal probability distributions** is denoted by $\mathbb{P}_m$ for $m-1$ extremes, with $m-1 \geq 2$ (i.e. $m \geq 3$). The random structure function $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_m$ having the codomain $\mathbb{P}_m$ is given by

$$\mathcal{P}(\{d_Y\}) = f_{Y|\{d_Y\}}(y) = e^{\lambda_0+\lambda_1 y+\lambda_2 y^2+...+\lambda_m y^m},\ y \in \mathcal{X}_Y(\{d_Y\}) \quad (1.17)$$

such that $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)})$.

The principle says that, for the purpose of constructing a **multi-extremal** the probability distribution of Y, the exact available information must read: the knowledge of $\mathcal{X}_Y(\{d_Y\})$ and of $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}$.

1.8 The formulation of the targeted aim

The actual targeted aim of this dissertation is to develop **efficient software programs** by means of **numerical mathematical methods** to compute the following:

1. Minimum information **monotone** probability distributions of Y stated by (1.15), both in discrete and continuous cases of Y.

   i.e. $\lambda(d_Y) = (\lambda_1(d_Y))$ is computed with subject to the predetermined $d_Y = (\mu_Y^{(1)})$ as well as the predetermined $\mathcal{X}_Y(\{d_Y\})$

2. Minimum information **uni-extremal** probability distributions of Y stated by (1.16), both in discrete and continuous cases of Y.

   i.e. $\lambda(d_Y) = (\lambda_1(d_Y), \lambda_2(d_Y))$ is computed with subject to the predetermined $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$ as well as the predetermined $\mathcal{X}_Y(\{d_Y\})$

and by using the **object oriented programming concept**. The computed monotone as well as the uni- extremal probability distributions of Y are used for the development of certain **stochastic procedures**, especially the **prediction procedures**.

In other words, the computations of **random structure functions** $\mathcal{P}(\{d_Y\})$ in both **monotone** and **uni- extremal** cases in form of $\lambda(d_Y)$ defined by (1.14) for $m \in \{1, 2\}$ is the targeted aim of this thesis.

A practical situation demanding the necessity of a prediction procedure, namely the computation of an **optimal prediction interval**, has already been exemplified right at the beginning of this thesis work.

My programming work regards the programming of numerical mathematical methods for solving systems of simultaneous equations. The well known numerical solution procedure for solving systems of simultaneous equations is Newton Raphson procedure. **Usage of the Newton Raphson procedure necessitates extreme skillful programming techniques. In other words, if the programmer is not skilled enough to use the Newton Raphson procedure, he must avoid using it, otherwise he shall do more harm than good. As a skilled programmer, I have used the Newton Raphson procedure to big successes and it had meant a backbreaking amount of programming work for me. The skilful programming for to make the Newton Raphson procedure find the numerical solutions is a significantly major part of this dissertation**.

The usage of multi- extremal probability distributions of Y stated by (1.17) have practically **no relevance** in the field of stochastic science and therefore have no imminent importance in developing any stochastic procedure. Therefore, discussions about multi- extremal probability distributions are carried out briefly, i.e. without any elaborations or any practical implementations.

Of course, a sound **character analysis** of constant, monotone (in forms of either (1.15) or (1.16)) and uni- extremal (in form of (1.16)) probability distributions of Y are eminently important from the **stochastic point of view**. The probability distributions of Y of monotone types (in form exclusively of (1.15)) and of uni- extremal types (in form exclusively of (1.16))) are therefore termed as **standard** minimum information probability distributions, whose characteristic properties are intensively analyzed in the chapter 6. Trivially, a (standard) minimum information constant probability distribution of Y is nothing different from an usual constant probability distribution of Y.

This aforesaid character analysis is also an aim of my dissertation next to the aforesaid targeted aim of my dissertation, which principally deals with the role of the **first moment** or the **first two moments** in characterizing the types of the standard minimum information probability distributions.

Lastly, for the sake of **completeness** of my dissertation, we should justify the usage of the exponential polynomial probability distributions of Y (stated in

(1.2) and (1.3)) as approximating probability distributions. This justification necessitates the discussions or presentations of certain essential properties of the exponential polynomial probability distributions.

One of the most important properties of an exponential polynomial probability distribution of Y (referred to the representations (1.2) and (1.3)) for any $m \in \mathbb{N}$ is its unique existence with subject to the provided available information in form of d_Y and $\mathcal{X}_Y(\{d_Y\})$ (elaborated in the sections 4.3 and 4.4). Other important properties of the same are discussed too.

In course of our discussions about the general properties of exponential polynomial probability distributions of Y, we shall use the notation $\mathbb{P}$ giving the family of all the exponential polynomial distributions of Y.

Before we draw this chapter to a close, let us mention certain important points to be noted:

- Since d_Y determines $\lambda(d_Y)$ uniquely, d_Y and $\lambda(d_Y)$ are **equivalent** to each other and the random structure function $\mathcal{P}(d_Y)$ is therefore a function of d_Y or equivalently of $\lambda(d_Y)$.

- Because of the very fact that d_Y determines a probability distribution of Y uniquely, for the sake of higher degree of clarity, we could think of writing Y as $Y|\{d_Y\}$. The same is the reason, for which we could think of writing $\mathcal{X}_Y(\{d_Y\})$ instead of simply $\mathcal{X}_Y$.

  However, if the notations are completely unambiguous, for the sake of simplicity, we shall use the notations Y and $\mathcal{X}_Y$.

- Generally, an empirically estimated value of the moment $\mu_Y^{(i)} = E[Y^i]$, $i \in \{1, 2, \ldots, m\}$ has to be denoted by $\widehat{\mu}_Y^{(i)}$. In the same way, an empirically estimated value of the variance of Y denoted by $\sigma_Y^2 = E[(Y - \mu_Y^{(1)})^2]$ has to be principally denoted by $\widehat{\sigma}_Y^{(1)}$. However, for the sake of simplicity, we shall omit the hat symbol $\widehat{\ }$ in course of our discussions, because the meanings in individual cases, which are discussed in this thesis, are completely unambiguous.

Referring to the skillful usage of the Newton Raphson procedure, let me give an important statement pertaining to what has been possibly performed by a particular group of three authors. As a matter of fact, the authors of

[31] made only an effort in making use of the Newton Raphson procedure, which has been termed by them as a hybrid approach. This programming effort is rather **wobble** and is noway near the proper skillful usage of the Newton Raphson procedure. This wobbliness is rather grave, because the authors of [31] (in accordance with their works published in their joint paper) have completely ignored giving the percentage volume of the input space of moments that at all takes care of delivering the program-outputs. Not only this, they have completely ignored taking care of the following basics of programming techniques:

1. The running problems of the software-programm arising out of overflow or underflow errors.

2. The convergence of the Newton Raphson procedure and the conditions for the convergence of the same.

As the immediate subsequent step, we shall discuss the third element of the Bernoulli space, namely the **random structure function**.

Chapter 2

The random structure function $\mathcal{P}$

2.1 The random variable Y

In probability theory, the random variables are defined as measurable functions on Ω and three types of random variables are distinguished:

1. **Type 1**: Random variable Y of discrete type, characterized by the fact, that the range of variability $\mathcal{X}_Y$ is denumerable. Equivalently, the distribution function is a step function.

2. **Type 2**: Random variables Y of continuous type, characterized by the fact, that the range of variability $\mathcal{X}_Y$ is non-denumerable and the distribution function has the following representation:

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt \tag{2.1}$$

   Equivalently, the distribution function is absolutely continuous.

3. **Type 3**: Random variables Y of degenerated type, characterized by the fact, that the range of variability $\mathcal{X}_Y$ is non-denumerable, but the distribution function is not absolutely continuous.

However, our discussions shall be principally confined to the types 1 and 2 only.

2.2 The random structure function

For analyzing a situation described by a pair of variables (Y, d_Y) the selection of the following are necessary at the very least:

- The parameter space $\mathcal{D}_Y$, which contains all the potential empirical values of d_Y that cannot be excluded.
- The range of variability $\mathcal{X}_Y(\{d_Y\})$ of Y, which is imaged by every singleton subset of $\mathcal{D}_Y$ denoted by $\{d_Y\}$.

The selection of the random structure function $\mathcal{P}$ is important for **studying** the **randomness** of Y. The selection must be based on some knowledge about $\mathcal{D}_Y$ having an influence on the random structure of Y or more precisely said, about the underlying process that produces the values of Y in form of outcomes.

Though $\mathcal{P}$ can be well defined on a suitably chosen system of subsets of $\mathcal{D}_Y$ denoted by $\mathcal{T}_{D_Y}(\mathcal{D}_Y)$, as we have already mentioned, we shall confine our discussions to singleton subsets of $\mathcal{D}_Y$ denoted by $\{d_Y\} \subset \mathcal{D}_Y$ only, i.e. $\mathcal{T}_{D_Y}(\mathcal{D}_Y) = \{\{d_Y\} | d_Y \in \mathcal{D}_Y\}$. This singleton subset $\{d_Y\}$ can also be termed as an **initial condition** (or a **boundary condition**). So, we restate here, the random structure function $\mathcal{P}$ has $\mathcal{T}_{D_Y}(\mathcal{D}_Y)$ as its **domain of definition** and a particular family of the probability distributions of Y denoted by $\mathbb{P}_m$ as its **codomain**. We shall elaborate the standard cases of $\mathbb{P}_m$ for $m \in \{0, 1, 2\}$ in this chapter itself.

Definition 2.2.1 (Symbolic notation of the random structure function particularly for a discrete Y). *If Y happens to be **discrete**, corresponding to a given argument $\{d_Y\}$, the image of $\mathcal{P}$ described by the notation $\mathcal{P}(\{d_Y\})$ for $d_Y \in \mathcal{D}_Y$ is a discrete probability measure:*

$$\mathcal{P}(\{d_Y\}) = P_{Y|\{d_Y\}} \tag{2.2}$$

where any probability measure $P_{Y|\{d_Y\}}$ is simply defined by the corresponding probability mass function $f_{Y|\{d_Y\}}$:

$$f_{Y|\{d_Y\}}(y) = P_{Y|\{d_Y\}}(\{y\}), \quad y \in \mathcal{X}_Y(\{d_Y\}) \tag{2.3}$$

*such that the range of variability $\mathcal{X}_Y(\{d_Y\})$ of Y is a **set of finite number of discrete elements** only.*

Definition 2.2.2 (Symbolic notation of the random structure function particularly for a continuous Y). *If Y is **continuous**, the above relation (2.2) can be simply redefined by the probability density $f_{Y|\{d_Y\}}$ as*

$$\mathcal{P}(\{d_Y\}) = f_{Y|\{d_Y\}} \tag{2.4}$$

and (2.3) simply needs to be remodified as

$$f_{Y|\{d_Y\}}(y), \quad y \in \mathcal{X}_Y(\{d_Y\}) \tag{2.5}$$

*such that the variability range $\mathcal{X}_Y(\{d_Y\})$ is simply a **closed and bounded interval**.*

From now on, for the sake of simplicity, let us restate the following for our future references: the notation $f_{Y|\{d_Y\}}(y)$ shall stand for both **probability mass function** (if Y happens to be **discrete**) and **probability density function** (if Y happens to be **continuous**). But, the notation $P_{Y|\{d_Y\}}(\{y\})$ shall stand exclusively for a **probability mass function** of a discrete Y.

With reference to [54], the set of all probability functions is divided into a set of disjoint families of probability describing functions of Y. (A **probability describing function** stands for either a **probability mass function** or a **probability density function**).

Thus, selecting an appropriate random structure function is equivalent to the selection of an appropriate probability describing function for Y with subject to a given d_Y. This task will be done in two steps. Firstly, a suitable family of probability describing functions is selected and secondly a suitable family member is determined.

2.3 Partitioning of the set of probability describing functions

In the following, the **discrete probability distributions with finite supports** and the **continuous probability distributions with closed and bounded supports** will be investigated. Both these discrete and the continuous types, being symbolized by $f_{Y|\{d_Y\}}(y)$, are the functions of y for a given choice of d_Y.

Definition 2.3.1 (Basic difference between the probability distributions of discrete and of continuous types). *Both these discrete and the continuous types are basically distinguished by their* ***supports*** *and in fact,*

- *in discrete cases,*

  the support is given by $\mathcal{X}_Y(\{d_Y\}) = \{y_1, y_2, \ldots, y_N\}$. *Without any loss of generality, we assume* $y_1 < y_2 < \ldots < y_N$ *and denote the probability mass function of* Y *by* $f_{Y|\{d_Y\}}(y) = P_{Y|\{d_Y\}}(\{y_j\}) > 0, j = 1, 2, \ldots, N$.

- *in continuous cases,*

  the support is given by $\mathcal{X}_Y(\{d_Y\}) = \{y| -\infty < a \leq y \leq b < \infty\}$ *and denote the probability density function of* Y *by* $f_{Y|\{d_Y\}}(y) > 0$.

Definition 2.3.2 (Partitioning of all the probability distributions). $\mathbb{P}$ *being the set of* ***all*** *probability distributions of* Y *with* ***compact supports*** *(both in cases for discrete and continuous), referring to page 166 of [54], the following partition in disjoint families is considered:*

- $\mathbb{P}_0$ *is the set of all probability distributions, which represents uniform (or constant) probability distributions.*

- $$\mathbb{P} \setminus \mathbb{P}_0 = \bigcup_{k=1}^{\infty} \mathbb{P}_k \tag{2.6}$$

  where $\mathbb{P}_k$ *is the set of all probability distributions with exactly* $k-1$ *relative extremal points of the function* $f_{Y|\{d_Y\}}(y)$ *and all the elements of this set are qualitatively of the same type.*

Remark 2.3.1 (The mode of partitioning). *The above mode of partitioning says that, the* $\mathbb{P}$ *represents the set of* ***uniform*** *probability distributions,* ***monotonic*** *probability distributions (i.e.* $k = 1$*) as well as the probability distributions with* ***one*** *or* ***more relative extremal points*** *(i.e.* $k > 1$*) of* $f_{Y|\{d_Y\}}(y)$*.*

Remark 2.3.2 (Disjoint partitioning). *Clearly, (2.6) represents a disjoint partition of the set of probability distributions. Moreover, the classes* $\mathbb{P}_0$*,* $\mathbb{P}_1$*,* $\mathbb{P}_2$*, ... are ordered according to the complexity of the the random structure, where the complexity is measured by the value of* k*:* ***higher*** *is the value of* k*,* ***larger*** *is the* ***complexity of the random structure****.*

2.4 Efficiency of the exponential polynomial distribution

As we have already mentioned, a suitably chosen exponential polynomial probability distribution has to be designed to approximate the **unknown but existing probability distribution in a given situation**. It is therefore extremely important for us to know, whether the exponential polynomial probability distribution is efficient enough to approximate this unknown but existing probability distribution. In plain and simple words, the word **efficiency** in this regard is ascribed to the ability of the exponential polynomial probability distribution to approximate this unknown but existing probability distribution in the situation.

This efficiency adds to the justification of using the **exponential polynomial probability distribution** as the **random structure function**.

This situation oriented need based probability distribution of Y described by the probability describing function $f_Y(y)$, $y \in \mathcal{X}_Y$ (as stated in (1.1)) may be in the discrete form or even in the continuous form.

We shall show in this very section, that the exponential polynomial distribution approximates both the discrete and the continuous form of this situational oriented need based probability distribution well enough. In fact, the exponential polynomial distribution **coincides with the discrete form completely** and **approximates the continuous form to any desired predetermined degree of accuracy**.

Realistically speaking, in both discrete and continuous cases of Y, we do not have any reason to assume that $f_Y(y) = 0$ for some $y \in \mathcal{X}_Y$. In fact, the zero values of $f_Y(y)$ are of no interest at all and therefore can be kept well out of consideration. Moreover, in reality, we do not have any reason to assume that $f_Y(y)$ could be discontinuous (in continuous cases of Y) either. So, keeping this in mind, we shall proceed.

With subject to the following conditional restrictions:

- If $f_Y(y)$ describes a probability mass function defined on a **finite set** $\mathcal{X}_Y$ of values of y, the description of $f_Y(y)$ is restricted to $f_Y(y) > 0$ without violation of any basic rule for describing a probability distribution
- If $f_Y(y)$ describes a probability density function defined on a **compact interval** $\mathcal{X}_Y$, the description of $f_Y(y)$ is restricted to its **continuity** as well as $f_Y(y) > 0$ throughout that interval

with regard to the page 163 of [54], $f_Y(y)$ is **determinable** or **approximatively expressible** by a certain **finite number of moments**. Exactly in this regard, $f_Y(y)$ is **expressible** in the following form:

$$f_Y(y) \approxeq e^{\sum_{i=0}^{n} \lambda_i y^i} \tag{2.7}$$

such that

- in the **discrete case**, $n = N - 1$, where N is the number of elements of the set of values of y and the representation is **exact**
- in the **continuous case**, the representation can be made **arbitrarily accurate** by adjusting the value of n

For both discrete and continuous cases of Y, since for every $f_Y(y) > 0$, the logarithmic expression $\log(f_Y(y))$ is always a real function of y, we can always rewrite the expression of $f_Y(y)$ in that case in the following way

$$\begin{aligned} f_Y(y) &= e^{\log(f_Y(y))} \\ &= e^{g(y)} \end{aligned} \tag{2.8}$$

Now, let us discuss the **discrete** and the **continuous** cases one by one in the subsequent subsections.

2.4.1 Exact representation of $f_Y(y)$ for the discrete case

Proposition 2.4.1. *Let $\mathcal{X}_Y = \{y_1, y_2, \ldots, y_N\}$, where $y_1 < y_2 < \ldots < y_N$.*

*In this discrete case of $f_Y(y)$, we shall go **only** by $f_Y(y) > 0$.*

Then, in that case, $f_Y(y)$, $y \in \mathcal{X}_Y$ is representable as

$$f_Y(y_j) = e^{\sum_{i=0}^{N-1} \lambda_i y_j^i}, \quad j = 1, 2, \ldots, N \tag{2.9}$$

*Proof of the **proposition 2.4.1**.* By using (2.8), we get a series of N real values as follows:

$$g(y_j) = \log\left(f_Y(y_j)\right), \quad j = 1, 2, \ldots, N \tag{2.10}$$

with the help of which, we can always construct a polynomial of degree $N-1$ given by

$$g(y_j) = \sum_{i=0}^{N-1} \lambda_i y_j^i, \quad j = 1, 2, \ldots, N \tag{2.11}$$

where the coefficients λ_i, $i = 0, 1, 2, \ldots, N-1$ of the integral powers of y_j, for $j = 1, 2, \ldots, N$ respectively are uniquely determined by solving the system of N simultaneous linear equations (2.11) in $\lambda_0, \lambda_1, \lambda_2, \ldots, \lambda_{N-1}$ with subject to the known values of $g(y_j)$, $j = 1, 2, \ldots, N$ by **inverting the non-singular Vandermonde matrix** stated immediately below by the following working rule:

$$\begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{N-1} \end{pmatrix} = \begin{pmatrix} 1 & y_1 & y_1^2 & \ldots & y_1^{N-1} \\ 1 & y_2 & y_2^2 & \ldots & y_2^{N-1} \\ 1 & y_3 & y_3^2 & \ldots & y_3^{N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & y_N & y_N^2 & \ldots & y_N^{N-1} \end{pmatrix}^{-1} \begin{pmatrix} g(y_1) \\ g(y_2) \\ g(y_3) \\ \vdots \\ g(y_N) \end{pmatrix} \tag{2.12}$$

Therefore, by combining (2.8) (2.10) and (2.11), we get the uniquely determined representation of $f_Y(y)$ as

$$f_Y(y_j) = e^{\sum_{i=0}^{N-1} \lambda_i y_j^i}, \quad j = 1, 2, \ldots, N \tag{2.13}$$

and hence our **proposition 2.4.1** gets proven. □

Remark 2.4.1 (Representation for the discrete case is exact). *The probability mass function* $e^{\sum_{i=0}^{N-1} \lambda_i y_j^i}$*, which represents the* ***existing*** *but* ***unknown*** *probability mass function* $f_Y(y_j)$*, is the* ***exact*** *representation (in the discrete case) of* $f_Y(y_j)$*.*

2.4.2 Approximative representation of $f_Y(y)$ for the continuous case

At first, let us state the **Weierstrass's theorem on approximations by polynomials**.

Theorem 2.4.1 (Weierstrass's theorem on approximations by polynomials). *Every continuous function* $g(y)$ *defined in a closed and bounded interval* $[a, b]$ *can be* ***uniformly approximated*** *in* $[a, b]$ *by a polynomial. That is, for any arbitrarily preassigned positive number* $\epsilon > 0$*, there exists a polynomial* $P_\epsilon(y)$*, such that* $|g(y) - P_\epsilon(y)| < \epsilon$ *for every* $y \in [a, b]$*.*

The formal analytical proof of the **theorem 2.4.1** by means of the usage of the **Fourier series** is well **stated** and **proved** in the pages 446 - 448 of [32], §14.08 - 14.081.

Remark 2.4.2. *In the year 1885, Weierstrass has proven* ***two*** *important theorems regarding the analytic representability of univariate functions, the* ***first theorem*** *of Weierstrass being exactly the aforesaid* ***theorem 2.4.1*** *and is well stated in the chapter of* ***interpolation*** *on the page 47 of [36]. The idea of this very analytic representability is referred to the historic publication [58] of Weierstrass (in German language).*

The very justification for replacing a given (univariate) function by a ***polynomial*** *or by a* ***finite trigonometric series*** *rests on these* ***two theorems*** *proved by Weierstrass.*

Remark 2.4.3 (Brief description of the statement of the theorem 2.4.1). *Let the degree of the polynomial* $P_\epsilon(y)$ *(in* y*) be denoted by the* ***natural number*** $N(\epsilon)$*. This natural number* $N(\epsilon)$ *depends on the* ***predeterminately arbitrarily chosen*** *small positive number* ϵ (> 0)*.*

The expression of the approximating polynomial $P_\epsilon(y)$ *of degree* $N(\epsilon)$*, where the* ***coefficients*** *of* y^i *are denoted by* λ_i *for* $i \in \{0, 1, 2, \ldots, N(\epsilon)\}$*, is there-*

fore defined as

$$P_\epsilon(y) = \sum_{i=0}^{N(\epsilon)} \lambda_i y^i \tag{2.14}$$

So, the theorem 2.4.1 says that for any ***arbitrarily predeterminately*** *chosen* ***small positive number*** ϵ (> 0)*, there exists a* ***natural number*** $N(\epsilon)$*, such that*

$$|g(y) - P_\epsilon(y)| < \epsilon \text{ for every } y \in [a, b]$$

and that is, by (2.14),

$$\left| g(y) - \sum_{i=0}^{N(\epsilon)} \lambda_i y^i \right| < \epsilon \text{ for every } y \in [a, b] \tag{2.15}$$

Remark 2.4.4. ***Notably****, the real valued coefficients* λ_i *of* y^i*, such that* $i \in \{0, 1, 2, \ldots, N(\epsilon)\}$ ***depend*** *on the following:*

- *the predeterminately assigned value of* ϵ *and subsequently on the natural number* $N(\epsilon)$*.*
- *the choice of the end points of the closed and the bounded interval, namely* a *and* b *(such that* $a < b$*).*
- *the behavior of* $g(y)$*, i.e. how often* $g(y)$ *fluctuates within* $y \in [a, b]$*.*

Remark 2.4.5. *The* ***goodness*** *of the approximation of* $g(y)$ *by the approximating polynomial* $P_\epsilon(y)$ *is* ***understandably*** *given by the* ***smallness*** *of the predetermined* ϵ*.*

With this, we proceed to prove our principally targeted proposition.

Proposition 2.4.2. *Let* $\mathcal{X}_Y = \{y \mid a \leq y \leq b\}$*, such that* $a < b$*.*

In this continuous case of $f_Y(y)$*, we shall go by the* ***continuity*** *(and thereby the* ***boundedness****) of the density function (i.e. of* $f_Y(y)$*) as well as by the* ***strict positivity*** $f_Y(y) > 0$ *throughout the interval* $[a, b]$*.*

Then, in that case, $f_Y(y)$*,* $y \in \mathcal{X}_Y$ *is approximatively representable as follows:*

$$f_Y(y) \approx e^{\sum_{i=0}^{n} \lambda_i y^i}, y \in [a, b] \tag{2.16}$$

where the ***natural number*** n *is determined by the predeterminately chosen degree of accuracy of the aforesaid approximation.*

Proof of the ***proposition 2.4.2****.* In order to justify the replacement of the function $g(y)$, $a \leq y \leq b$ stated in (2.8) by a polynomial, we need to refer to the **Weierstrass's theorem 2.4.1**.

Accordingly, since $g(y) = \log\left(f_Y(y)\right)$ is **continuous** in $[a, b]$, for every arbitrarily preassigned $\epsilon > 0$, there exists a positive integer $N(\epsilon)$, such that

$$\left| g(y) - \sum_{i=0}^{N(\epsilon)} \lambda_i y^i \right| < \epsilon \tag{2.17}$$

Thus, (2.17) necessarily implies that

$$f_Y(y) = e^{g(y)} \approx e^{\sum\limits_{i=0}^{N(\epsilon)} \lambda_i y^i} \tag{2.18}$$

and thereby the very fact that $e^{\sum\limits_{i=0}^{N(\epsilon)} \lambda_i y^i}$ can approximate $f_Y(y)$ to **any desired level of accuracy**. This level of accuracy is determined by the value of ϵ.

This proves our **proposition 2.4.2**. □

Remark 2.4.6 (Representation for the continuous case is well approximated). *The probability density function* $e^{\sum\limits_{i=0}^{N(\epsilon)} \lambda_i y^i}$*, which can* ***approximate*** *the* ***existing*** *but* ***unknown*** *probability density function* $f_Y(y)$ *to any desired degree of accuracy (in the continuous case), is inexact to a predetermined degree of smallness. In plain words, this degree of smallness can be* ***arbitrarily*** *chosen.*

Remark 2.4.7 (The natural number $N(\epsilon)$). *The reader of this dissertation may say that the natural number* $N(\epsilon)$ *is an* ***optimal choice*** *of the number of moments needed to approximate the probability density* $f_Y(y)$, $y \in [a, b]$ ***optimally****.*

By (2.18), it is clear that from the ***mathematical point of view*** *the probability density* $f_Y(y)$, $y \in [a, b]$ *is well approximated by the probability density* $e^{\sum\limits_{i=0}^{N(\epsilon)} \lambda_i y^i}$, $y \in [a, b]$*. This very mathematical point of view serves* ***purely*** *as the* ***justification*** *of this aforesaid way of approximation.*

But, from the ***statistical point of view****, as already discussed,* $f_Y(y)$ *is* ***generally unknown****. Therefore, because of this very unknownness, the*

*predetermination of ϵ is **not possible** and hence there is **no question** of the determination of $N(\epsilon)$. So, in this very regard (i.e. with regard to the **statistical point of view**), the question of $N(\epsilon)$ being acceptable as the aforesaid optimal choice **is rather vague**.*

*However, from the statistical point of view, the **moments** of the probability distribution of Y given by the probability density $f_Y(y)$, $y \in [a, b]$ may be correctly estimable by means of good estimation procedures. In this very regard, as a brief note, the aforesaid optimal choice is given by consideration of certain characteristic properties of **these moments**. These moments do fulfill certain **optimality conditions**, the formal discussions of which are given in the coming **chapter 5**).*

Remark 2.4.8 (The important role of the natural number $N(\epsilon)$). *It is only intuitively assertible that the $e^{\sum_{i=0}^{N(\epsilon)} \lambda_i y^i}$, $y \in [a, b]$ is a **consistent density estimator** of $f_Y(y)$, $y \in [a, b]$ (see the **proposition 3.4.2** of the **subsection 3.4.4** for the formal treatment.). However, it has **not** been specifically proved here that $N(\epsilon)$ **increases** strictly monotonically with the strict monotonic **decrease** in ϵ (qualitatively speaking, it remains to be formally proved that the **largeness** of $N(\epsilon)$ is **solely** directly proportional to the **smallness** of ϵ).*

2.4.3 Short summary of discrete and continuous representations

Purely for the sake of clarity, let us give a short summary of both the **discrete** and **continuous** representations of $f_Y(y)$, $y \in \mathcal{X}_Y$.

If Y happens to be **discrete**, we know that the **exact representation** of $f_Y(y)$, $y \in \mathcal{X}_Y$ is $e^{\sum_{i=0}^{N-1} \lambda_i^{(D)} y_j^i}$, $\mathcal{X}_Y = \{y_1, y_2, \ldots, y_N\}$. Here, $\lambda_0^{(D)}$ is uniquely determinable by the values of $\lambda_i^{(D)}$ with $i \in \{1, 2, \ldots, N-1\}$.

By taking $m = N - 1$, the representation of $f_Y(y)$ for a discrete Y can be rewritten as $e^{\sum_{i=0}^{m} \lambda_i^{(D)} y_j^i}$. Here, m is the number of key moments of the probability distribution. Therefore, the representation of $f_Y(y)$, $y \in \mathcal{X}_Y$ in

the discrete case is restated as:

$$e^{\sum_{i=0}^{m} \lambda_i^{(D)} y_j^i}, \quad j \in \{1, 2, \ldots, N\} \tag{2.19}$$

where, $\lambda_0^{(D)}$ being uniquely determinable by the values of $\lambda_i^{(D)}$ with $i \in \{1, 2, \ldots, m\}$.

The first m moments of the probability distribution (2.19) shall be usually denoted by $\mu_Y^{(i,D)}$, $i \in \{1, 2, \ldots, m\}$.

If Y happens to be **continuous**, we know that the **approximative representation** of $f_Y(y)$, $y \in \mathcal{X}_Y$ is $e^{\sum_{i=0}^{m} \lambda_i^{(C)} y^i}$, $\mathcal{X}_Y = [a, b]$. Here, $\lambda_0^{(C)}$ is uniquely determinable by the values of $\lambda_i^{(C)}$ with $i \in \{1, 2, \ldots, m\}$.

Therefore, the representation of $f_Y(y)$, $y \in \mathcal{X}_Y$ in the continuous case is restated as:

$$e^{\sum_{i=0}^{m} \lambda_i^{(C)} y^i}, \quad a \leq y \leq b \tag{2.20}$$

The first m moments of the probability distribution (2.20) shall be usually denoted by $\mu_Y^{(i,C)}$, $i \in \{1, 2, \ldots, m\}$.

2.4.4 Continuous case as the case of approximation

We have seen that the probability distribution of Y represented by **either** a probability mass function in case of a discrete Y **or** by a probability density function in case of a continuous Y can be represented by a exponential polynomial function. As we know, in a given situation, this probability distribution of Y denoted by $f_Y(y)$ is **existing**, but **generally unknown**.

In this subsection, we shall briefly discuss, in which cases, the existing and (unknown) probability distribution of Y can be chosen to be **continuous**.

In plain words, this is precisely to say that, if the number of elements of the support of the **discrete** probability distribution of Y is **large enough**, the probability distribution of Y may be chosen to be a **continuous** one. In this regard, the **continuous case** of Y is regarded as the **case of approximation** of the **discrete case** of Y. Keeping this in mind, we shall proceed.

The **number of extremal points** of the probability distribution of Y represented either by its probability mass function (if Y is discrete) or by its probability density function (if Y is continuous) is the **basic characteristic property** for the choice of the type of the probability distribution. With reference to the subsequently stated **definition 3.3.1** (of the **subsection 3.3.4**), this **characteristic property** is basically described by the λ_i ($i \in \{1, 2, \ldots, m\}$) values. In this very regard, these λ_i values must **remain the same** for both the discrete and continuous cases of Y.

So, referring to the **subsequently shown result** (3.95) belonging to the **proposition 3.3.3** (of the **subsection 3.3.4**), for any fixedly chosen m, such that $m \leq N - 1$, by taking

- $y_1 = a$ and $y_N = b$
- $\lambda_i^{(D)} = \lambda_i^{(C)}$, $i \in \{1, 2, \ldots, m\}$

to be fixed, we arrived at $\lim\limits_{N \to \infty} \mu_Y^{(i,D)} = \mu_Y^{(i,C)}$ for $i \in \{1, 2, \ldots, m\}$.

Even if m is not fixedly chosen, but is allowed to increase with N with subject to $m = N-1$, we can establish the following statement: for a sufficiently large value of N, almost all the moments of both the discrete and the continuous probability distributions tend to coincide with each other.

Exactly, the other way around, that is **conversely**, if we fix

- $y_1 = a$ and $y_N = b$
- $\mu_Y^{(i,D)} = \mu_Y^{(i,C)}$, $i \in \{1, 2, \ldots, m\}$

then we put the following intuitively clear statement: $\lambda_i^{(D)} \to \lambda_i^{(C)}$ as $N \to \infty$ for every $i \in \{1, 2, \ldots, m\}$ (this is also referred to the given **statement 3.3.2** of the **subsection 3.3.4**). In this sense, $\lambda_i^{(C)}$ is the **limiting value** of $\lambda_i^{(D)}$ for every $i \in \{1, 2, \ldots, m\}$.

Thus, in this very **limiting sense**, the probability distribution described by the probability density function $e^{\sum\limits_{i=0}^{m} \lambda_i^{(C)} y^i}$, $a \leq y \leq b$ is the **approximated continuous probability distribution** of the probability distribution described by the probability mass function $e^{\sum\limits_{i=0}^{m} \lambda_i^{(D)} y_j^i}$, $j \in \{1, 2, \ldots, N\}$.

Conclusively, the **exact representation** of the discrete probability distribution described by the probability mass function $e^{\sum_{i=0}^{m} \lambda_i^{(D)} y_j^i}$, $j \in \{1, 2, \ldots, N\}$, the exactness being described by $m = N - 1$, can be alternatively well represented by the aforesaid approximated continuous probability distribution, **provided** N is sufficiently or at least reasonably high.

Moreover, cases may arise, when it is not possible to take $m = N - 1$ for **practical reasons**, but we need to restrict the value of m to a certain value with subject to $m < N - 1$. In such cases, the representation of $f_Y(y)$ in the discrete case however may become inexact to a small degree. This means nothing but the very that, the probability mass function given by (2.19), namely $e^{\sum_{i=0}^{m} \lambda_i^{(D)} y_j^i}$, $j \in \{1, 2, \ldots, N\}$ becomes **theoretically inexact** to a small degree.

In such cases (of $m < N-1$) **nothing changes** or anything is even expected to change, except the aforesaid fact that the representation may become inexact to a certain negligible degree, **provided** the most important aforesaid **characteristic property** is kept unchanged after the value of m being conveniently reduced. This is to say that, even in such cases, if N is sufficiently high, the aforesaid **approximated continuous probability distribution** remains good usable as the **replacement** of the **discrete probability distribution** in the sense that, as already stated above, corresponding to the fixedly chosen $\mu_Y^{(i,D)} = \mu_Y^{(i,C)}$, $i \in \{1, 2, \ldots, m\}$, we have $\lambda_i^{(D)} \to \lambda_i^{(C)}$ as $N \to \infty$ for every $i \in \{1, 2, \ldots, m\}$, but for a fixed value of m.

This very assertion shall be handled subsequently once again and shall be illustrated by two simple numerical examples in the chapter 3.

2.4.5 An informatory observation

We have seen that the approximative continuous probability distribution described by the probability density function $e^{\sum_{i=0}^{m} \lambda_i^{(C)} y^i}$, $a \leq y \leq b$ can be **a reasonably good alternative** to an exact discrete probability distribution described by the probability mass function $e^{\sum_{i=0}^{m} \lambda_i^{(D)} y_j^i}$, $j \in \{1, 2, \ldots, N\}$ with $m = N - 1$, provided N is reasonably large. For this, we have principally assumed that all the individual probabilities of the discrete probability dis-

tribution are **different from zeros**.

For the construction of this approximative continuous probability distribution, basically $\lambda_i^{(D)}$ ($i \in \{1, 2, \ldots, m\}$) are replaced by $\lambda_i^{(C)}$ ($i \in \{1, 2, \ldots, m\}$) with subject to the maintenance of $\mu_Y^{(i,D)} \approx \mu_Y^{(i,C)}$ ($i \in \{1, 2, \ldots, m\}$) so as to get the probability density function $e^{\sum_{i=0}^{m} \lambda_i^{(C)} y^i}$, $a \leq y \leq b$.

So, in this particular regard, there **cannot** be any reason to assume that the probability density function can either be **zero** or be **discontinuous** at any point within the interval $[a, b]$. In other words, this probability density function needs to be **non-zero as well as continuous** within the entire interval $[a, b]$.

2.4.6 Conclusive points

This approximative continuous probability distribution of Y given by its density $f_{Y|\{d_Y\}}(y) = e^{\sum_{i=0}^{m} \lambda_i^{(C)} y^i}$, $a \leq y \leq b$ ($f_{Y|\{d_Y\}}(y)$ being **uniquely determinable** by d_Y) has therefore certain well known **characteristic properties**, which are stated as follows:

1. $f_{Y|\{d_Y\}}(y) > 0$ for every $y \in [a, b]$.

(2.21)

2. $f_{Y|\{d_Y\}}(y)$ continuous and bounded in $[a, b]$.

(2.22)

3. $f_{Y|\{d_Y\}}(y)$ is derivable of any order in $[a, b]$.

4. $f_{Y|\{d_Y\}}(y)$ certainly **cannot have uncountably many relative extremal points** in $[a, b]$. This is precisely the idea, which we have used to **classify the different types of probability distributions in continuous cases previously**.

5. Even if m is fixedly chosen with subject to $m < N - 1$, but N is **reasonably large**, this very density $f_{Y|\{d_Y\}}(y)$, $y \in [a, b]$ **can be** made

to replace the discrete probability distribution described by $e^{\sum_{i=0}^{m} \lambda_i^{(D)} y_j^i}$, $j \in \{1, 2, \ldots, N\}$ comfortably.

For the construction of families of different types of probability distributions of Y, the fulfillment of the characteristic properties (1), (2), (3) and (4) are evident.

Among these characteristic properties, the characteristic properties (1) (i.e. (2.21)) and (2) (i.e. (2.22)) are of **primary importance**.

Moreover, we propose to state the following for the sake of our **future references**:

Statement 2.4.1 (Differentiation between the discrete and the continuous cases). *In course of our discussions, we shall normally use* λ_i, $i \in \{1, 2, \ldots, m\}$ *instead of* $\lambda_i^{(D)}$ *or* $\lambda_i^{(C)}$, *unless and until the differentiation between discrete and continuous cases is of absolute necessity.*

Statement 2.4.2 (Indexing of moments). *In course of our discussions, we shall use the indices* $\boldsymbol{m}$, $\boldsymbol{i}$ *and* $\boldsymbol{n}$ *for the following purposes:*

- *the index* $\boldsymbol{m}$ *stands for the exact number of moments to be used to construct the approximating probability distribution of* Y.
- *the index* $\boldsymbol{i}$ *stands* ***principally*** *for* ***ranking every moment*** *of* Y *starting from* $\boldsymbol{1}$ *to* $\boldsymbol{m}$, *i.e.* $i \in \{1, 2, \ldots, m\}$. *Certain* ***minor exceptions*** *are however present in this dissertation.*
- *the index* $\boldsymbol{n}$ *serves to set the rankings of the moments under consideration, but* ***not necessarily*** *being restricted by* $\boldsymbol{m}$. *For eg.: we shall use* $\boldsymbol{n} \in \mathbb{N}$, $\boldsymbol{i = 2\,n}$, $\boldsymbol{i = 2\,n + 1}$ *or even* $\boldsymbol{i = n}$. *That is,* $\boldsymbol{n}$ *stands for any natural number meant for proper indexation of a moment of* Y.

Before we draw our discussions of this section to a close, we need to mention one more important property of the well known theoretical beta distribution.

Remark 2.4.9 (The beta distribution). *If $f^B_{X|\{d\}}(x)$, $0 \leq x \leq 1$ denotes the probability density function of the well known beta-distribution, we are well aware of the following:*

- *$f^B_{X|\{d\}}(x) = 0$ both at $x = 0$ and $x = 1$, when the probability density $f^B_{X|\{d\}}(x)$ is **uni-modal** within $x \in [0, 1]$.*

  This violates the aforesaid characteristic property (1) given by (2.21), namely $f^B_{X|\{d\}}(x) > 0$ for every $x \in [0, 1]$.

- *$f^B_{X|\{d\}}(x)$ is infinitely discontinuous **either** at $x = 0$ **or** at $x = 1$ or **both**, when the probability density $f^B_{X|\{d\}}(x)$ is either **monotone** or **bathtub** within $x \in [0, 1]$.*

  This violates the aforesaid characteristic property (2) given by (2.22), namely $f^B_{X|\{d\}}(x)$ is continuous and bounded for every $x \in [0, 1]$.

*Therefore, **for this very reason, the beta distribution cannot be regarded as the approximative continuous probability distribution of any discrete probability distribution with non-zero individual probabilities**.*

However, it can be regarded as a theoretical probability distribution.

Now, we proceed to discuss the different families of probability distributions of Y.

2.5 The families of probability distributions

A partition of the set $\mathbb{P}$ of probability distributions in **disjoint families** is given in (2.6). Nextly, the different families must be characterized by means of easily verifiable properties. The families are defined by means of the probability mass or probability density function and, therefore, it makes sense to characterize the families by qualitative properties of these functions, which can be understood and verified without any specific expertise.

In general, we symbolize the **compact** support of the probability distribution as $\mathcal{X}_Y(\{d_Y\}) = \{y_1, y_2, \ldots, y_N\}$ in the **discrete case** and as $\mathcal{X}_Y(\{d_Y\}) = [a, b]$ in the **continuous case**.

The elements of $\mathbb{P}$ are the probability mass functions or probability density functions (according as Y is discrete or continuous) in form of **exponential polynomial functions only**. So, if Y is continuous, then its probability density

- is continuous in $[a, b]$
- and its derivatives of all orders exist in $[a, b]$

So, let us discuss the important families of probability distributions one by one:

2.5.1 The constant family: $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \to \mathbb{P}_0$

The members of the family $\mathbb{P}_0$ (stated in the page 167 of [54]) are characterized by a constant probability mass or a probability density function and, therefore, this family is called the *constant family*.

The constant family constitutes the simplest family of probability distributions. Nevertheless, it has been proved to be of great importance for the development of the probability theory. It was taken as an appropriate description for the structure of randomness in the case of a game of chance, which marked the beginning of probability theory. Each member $f_{Y|\{d_Y\}}$ of the constant family is solely determined by the range of variability $\mathcal{X}_Y(\{d_Y\})$.

Definition 2.5.1 (The constant family $\mathbb{P}_0$). *As we know, $\mathcal{X}_Y(\{d_Y\})$ is predeterminately given, the probability mass or density function can be im-*

mediately specified as:

$$f_{Y|\{d_Y\}}(y) = \frac{1}{|\mathcal{X}_Y(\{d_Y\})|} \text{ for } y \in \mathcal{X}_Y(\{d_Y\}) \tag{2.23}$$

where

$$|\mathcal{X}_Y(\{d_Y\})| = \begin{cases} N & \text{in the discrete case} \\ b-a & \text{in the continuous case} \end{cases} \tag{2.24}$$

Remark 2.5.1 (Representation of a constant probability distribution). *In the view of the representation (2.7), for any natural number n, we have:*

$$f_{Y|\{d_Y\}}(y) = e^{\sum_{i=0}^{n} \lambda_i(d_Y) y^i} = \begin{cases} \frac{1}{N} & \text{in the discrete case} \\ \\ \frac{1}{b-a} & \text{in the continuous case} \end{cases} \tag{2.25}$$

Remark 2.5.2 (λ_i values of a constant probability distribution). *From (2.25), we easily obtain:*

$$\begin{aligned} \lambda_i(d_Y) &= 0 && \text{for } i \in \{1, 2, \ldots, n\} \\ \lambda_0(d_Y) &= \begin{cases} -\log N & \text{for the discrete case} \\ -\log(b-a) & \text{for the continuous case} \end{cases} \end{aligned} \tag{2.26}$$

2.5.2 The monotone family: $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_1$

This is the family $\mathbb{P}_1$ (stated in the page 168 of [54]) of probability distributions with monotone probability mass or probability density functions, i.e., with boundary extremes and without a relative extreme. Obviously, there are two subfamilies to be considered, which shall be characterized in this subsection.

Moreover, in continuous cases of Y, **as already mentioned**, the following must be importantly stated: for every point $y \in [a, b]$, $f'_{Y|\{d_Y\}}(y) \neq 0$ must necessarily hold. This makes sure that, the endpoints $y = a$ and $y = b$ are **principally not** considered as extremal points.

Definition 2.5.2 (Monotone increasing subfamily of $\mathbb{P}_1$). *In the **discrete case**, the probability mass function giving the probability of occurrence of $\{y\} \subset \mathcal{X}_Y(\{d_Y\})$ is said to be monotonically increasing, if*

$$f_{Y|\{d_Y\}}(y_1) < f_{Y|\{d_Y\}}(y_N) \tag{2.27}$$
$$f_{Y|\{d_Y\}}(y_i) \leq f_{Y|\{d_Y\}}(y_{i+1}) \text{ for } i = 1, \ldots, N-1 \tag{2.28}$$

*In the **continuous case**, the probability density function describing the probability distribution of Y is said to be monotonically increasing, if for every $\mathbf{y_1}, \mathbf{y_2} \in [a, b]$,*

$$f_{Y|\{d_Y\}}(\mathbf{y_1}) \leq f_{Y|\{d_Y\}}(\mathbf{y_2}) \text{ whenever } \mathbf{y_1} < \mathbf{y_2} \tag{2.29}$$

together with

$$f_{Y|\{d_Y\}}(a) < f_{Y|\{d_Y\}}(b) \text{ and} \tag{2.30}$$
$$f'_{Y|\{d_Y\}}(y) > 0 \text{ for every } y \in [a, b] \tag{2.31}$$

Definition 2.5.3 (Monotone decreasing subfamily of $\mathbb{P}_1$). *In the **discrete case**, the probability mass function giving the probability of occurrence of $\{y\} \subset \mathcal{X}_Y(\{d_Y\})$ is said to be monotonically decreasing, if*

$$f_{Y|\{d_Y\}}(y_1) > f_{Y|\{d_Y\}}(y_N) \tag{2.32}$$
$$f_{Y|\{d_Y\}}(y_i) \geq f_{Y|\{d_Y\}}(y_{i+1}) \text{ for } i = 1, \ldots, N-1 \tag{2.33}$$

*In the **continuous case**, the probability density function describing the probability distribution of Y is said to be monotonically decreasing, if for every $\mathbf{y_1}, \mathbf{y_2} \in [a, b]$,*

$$f_{Y|\{d_Y\}}(\mathbf{y_1}) \geq f_{Y|\{d_Y\}}(\mathbf{y_2}) \text{ whenever } \mathbf{y_1} < \mathbf{y_2} \tag{2.34}$$

together with

$$f_{Y|\{d_Y\}}(a) > f_{Y|\{d_Y\}}(b) \text{ and} \tag{2.35}$$
$$f'_{Y|\{d_Y\}}(y) < 0 \text{ for every } y \in [a, b] \tag{2.36}$$

Importantly, in this case of the monotone family the range of variability $\mathcal{X}_Y(\{d_Y\})$ alone is not enough to determine the probability measure $P_{Y|\{d_Y\}}$ or the probability density $f_{Y|\{d_Y\}}$ (as the case may be) uniquely and, therefore, additional appropriate information must be available for selecting the probability measure or density. This information must be unambiguously specified for any given situation.

2.5.3 The uni-extremal family: $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_2$

This is the family $\mathbb{P}_2$ (stated in the page 169 of [54]) of probability distributions with **exactly one** relative extreme. Just as in the case of the monotone family, there are two distinct subfamilies to be considered. The first one has exactly one **relative maximum**, called the **uni-modal** family and the second has exactly one **relative minimum**, called the **bathtub** family.

Moreover, in uni-extremal **continuous** cases, **as already mentioned**, the following two statements must be duly importantly put:

Statement 2.5.1 (Clear difference between the uni-extremity and monotonicity). *If either of the following happens to hold*

1. $f^{'}_{Y|\{d_Y\}}(a) = 0$ *and* $f^{''}_{Y|\{d_Y\}}(a) \neq 0$

   or *in general, if the extreme point is at $y = a$, then, at $y = a$, all the* ***odd*** *order derivatives of $f_{Y|\{d_Y\}}(y)$ are individually zero and at least one* ***even*** *order derivative of $f_{Y|\{d_Y\}}(y)$ is nonzero*

2. $f^{'}_{Y|\{d_Y\}}(b) = 0$ *and* $f^{''}_{Y|\{d_Y\}}(b) \neq 0$

   or *in general, if the extreme point is at $y = b$, then, at $y = b$, all the* ***odd*** *order derivatives of $f_{Y|\{d_Y\}}(y)$ are individually zero and at least one* ***even*** *order derivative of $f_{Y|\{d_Y\}}(y)$ is nonzero*

then the probability distribution of Y given by the probability density $f_{Y|\{d_Y\}}(y)$ is ***principally not*** *considered to be a monotone probability distribution, as the extreme point lies within $[a, b]$.*

Statement 2.5.2 (Uniqueness of the extremal point). *There can be* ***exactly one point*** *$y = y_0 \in [a, b]$, at which all the odd order derivatives of $f_{Y|\{d_Y\}}(y)$ are individually zero and at least one even order derivative of the same is nonzero, for eg. $f^{'}_{Y|\{d_Y\}}(y_0) = 0$ and $f^{''}_{Y|\{d_Y\}}(y_0) \neq 0$.*

Now, we proceed to elaborately define the uni-modal and bathtub families as follows:

Definition 2.5.4 (Uni-modal subfamily of $\mathbb{P}_2$). *In the* ***discrete case****, the probability mass function $f_{Y|\{d_Y\}}(y)$ giving the probability of occurrence of $\{y\} \subset \mathcal{X}_Y(\{d_Y\})$ is said to be uni-modal, if $f_{Y|\{d_Y\}}(y)$* ***increases first****, reaches its* ***maximum value at one point only****, say for $i = i_0$, such that $1 < i_0 < N$ and* ***decreases thereafter****, i.e.*

$$f_{Y|\{d_Y\}}(y_1) < f_{Y|\{d_Y\}}(y_{i_0}) > f_{Y|\{d_Y\}}(y_N) \tag{2.37}$$
$$f_{Y|\{d_Y\}}(y_i) \leq f_{Y|\{d_Y\}}(y_{i+1}) \text{ for } i = 1, 2, \ldots, i_0 - 1 \tag{2.38}$$
$$f_{Y|\{d_Y\}}(y_i) \geq f_{Y|\{d_Y\}}(y_{i+1}) \text{ for } i = i_0, i_0 + 1, \ldots, N - 1 \tag{2.39}$$

In the ***continuous case****, the probability density function $f_{Y|\{d_Y\}}(y)$ describing the probability distribution of Y is said to be uni-modal, if $f_{Y|\{d_Y\}}(y)$* ***increases first****, reaches its* ***maximum value at one point only****, say at $y = y_0$, such that $y_0 \in [a, b]$ and* ***decreases thereafter****. In that case, for every $\mathbf{y_1}, \mathbf{y_2} \in [a, b]$,*

$$f_{Y|\{d_Y\}}(a) \leq f_{Y|\{d_Y\}}(y_0) \text{ for } a \leq y_0 \tag{2.40}$$
$$f_{Y|\{d_Y\}}(y_0) \geq f_{Y|\{d_Y\}}(b) \text{ for } y_0 \leq b \tag{2.41}$$
$$f_{Y|\{d_Y\}}(\mathbf{y_1}) \leq f_{Y|\{d_Y\}}(\mathbf{y_2}) \text{ whenever } \mathbf{y_1} < \mathbf{y_2} \leq y_0 \tag{2.42}$$
$$f_{Y|\{d_Y\}}(\mathbf{y_1}) \geq f_{Y|\{d_Y\}}(\mathbf{y_2}) \text{ whenever } y_0 \leq \mathbf{y_1} < \mathbf{y_2} \tag{2.43}$$
$$f_{Y|\{d_Y\}}(a) = f_{Y|\{d_Y\}}(b) \Rightarrow f_{Y|\{d_Y\}}(a) < f_{Y|\{d_Y\}}(y_0) > f_{Y|\{d_Y\}}(b) \tag{2.44}$$

Importantly*, we should take care of the validity of the possible*

- ***maximality*** *at $y = y_0 = a$ represented by $f'_{Y|\{d_Y\}}(a) = 0$*

- ***maximality*** *at $y = y_0 = b$ represented by $f'_{Y|\{d_Y\}}(b) = 0$*

Definition 2.5.5 (Uni-bathtub subfamily of $\mathbb{P}_2$). *In the* ***discrete case****, the probability mass function $f_{Y|\{d_Y\}}(y)$ giving the probability of occurrence of $\{y\} \subset \mathcal{X}_Y(\{d_Y\})$ is said to be of bathtub shaped, if $f_{Y|\{y\}}(y)$* ***decreases first****, reaches its* ***minimum value at one point only****, say for $i = i_0$, such that $1 < i_0 < N$ and* ***increases thereafter****, i.e.*

$$f_{Y|\{d_Y\}}(y_1) > f_{Y|\{d_Y\}}(y_{i_0}) < f_{Y|\{d_Y\}}(y_N) \tag{2.45}$$
$$f_{Y|\{d_Y\}}(y_i) \geq f_{Y|\{d_Y\}}(y_{i+1}) \text{ for } i = 1, 2, \ldots, i_0 - 1 \tag{2.46}$$
$$f_{Y|\{d_Y\}}(y_i) \leq f_{Y|\{d_Y\}}(y_{i+1}) \text{ for } i = i_0, i_0 + 1, \ldots, N - 1 \tag{2.47}$$

In the ***continuous case****, the probability density function $f_{Y|\{d_Y\}}(y)$ describing the probability distribution of Y is said to be bathtub shaped, if $f_{Y|\{d_Y\}}(y)$* ***decreases first****, reaches its* ***minimum value at one point only****, say at $y = y_0$, such that $y_0 \in [a, b]$ and* ***increases thereafter****. In that case, for every $\mathbf{y_1}, \mathbf{y_2} \in [a, b]$,*

$$f_{Y|\{d_Y\}}(a) \geq f_{Y|\{d_Y\}}(y_0) \text{ for } a \leq y_0 \tag{2.48}$$
$$f_{Y|\{d_Y\}}(y_0) \leq f_{Y|\{d_Y\}}(b) \text{ for } y_0 \leq b \tag{2.49}$$
$$f_{Y|\{d_Y\}}(\mathbf{y_1}) \geq f_{Y|\{d_Y\}}(\mathbf{y_2}) \text{ whenever } \mathbf{y_1} < \mathbf{y_2} \leq y_0 \tag{2.50}$$
$$f_{Y|\{d_Y\}}(\mathbf{y_1}) \leq f_{Y|\{d_Y\}}(\mathbf{y_2}) \text{ whenever } y_0 \leq \mathbf{y_1} < \mathbf{y_2} \tag{2.51}$$
$$f_{Y|\{d_Y\}}(a) = f_{Y|\{d_Y\}}(b) \Rightarrow f_{Y|\{d_Y\}}(a) > f_{Y|\{d_Y\}}(y_0) < f_{Y|\{d_Y\}}(b) \tag{2.52}$$

Importantly*, we should take care of the validity of the possible*

- ***minimality*** *at $y = y_0 = a$ represented by $f'_{Y|\{d_Y\}}(a) = 0$*

- ***minimality*** *at $y = y_0 = b$ represented by $f'_{Y|\{d_Y\}}(b) = 0$*

Remark 2.5.3 (Increased complexity in uni-extremal cases). *In comparison to the monotone family, the uni-extremal family exhibits an* ***increased complexity*** *and, therefore, a larger amount of knowledge is needed for the selection of an appropriate probability distribution from the uni-extremal family than the same for the selection of a probability distribution from the monotone family.*

2.5.4 A multi-extremal family: $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_m$ with $m > 2$

The multi-extremal families (the family of type $\mathbb{P}_m$, $m > 2$ is stated in the page 171 of [54]) are characterized by two or more relative extreme points of the probability mass or density function. $m - 1$ gives the exact number of relative extremes. For each family, two subfamilies can be distinguished. The **first subfamily** is characterized by the fact that the first relative extreme is a **maximum**, while the extremes of the **second subfamily** start with a **minimum**.

Principally, only the families $\mathbb{P}_0$, $\mathbb{P}_1$ and $\mathbb{P}_2$ concern this thesis for detailed discussions, as they basically concern the science of stochastics. Therefore, the multi-extremal families shall not be discussed here.

Chapter 3

The minimum information and the maximum entropy

Before the computer code for the function $\lambda = \lambda(d_Y)$ is developed, the selection principle of **minimum information** is revisited by comparing it with another universal principle for modelling the randomness.

There is an interesting connection between the **minimum information principle** developed here and the **maximum entropy principle** (MEP) proposed by E. T. Jaynes in the year 1957 [24]. Before this relation can be established, the concept of *stochastic entropy* has been introduced.

Referring to the page 1 of [25], every probability distribution has some "uncertainty" associated with it. The stochastic entropy gives a quantitative measure of this uncertainty.

3.1 The stochastic entropy

The uncertainty of the occurrence an event with respect to a given random variable Y is described by the randomness, which in turn is described by the probability measure (or the probability density) of Y. This probability measure (or probability density) assigns to each possible event a particular probability.

The structure of the probability of an event with regard to the random variable Y is basically originated by d_Y.

Let us consider the probability measure $P_{Y|\{d_Y\}}(\{y\})$ for $y \in \mathcal{X}_Y(\{d_Y\}) = \{y_j \,|\, j = 1, 2, \ldots, N\}$. Then, the question arises, how to quantify the uncertainty associated with $P_{Y|\{d_Y\}}$, i.e. to find out the entropy of the probability distribution of Y described by $P_{Y|\{d_Y\}}$. This problem was solved by Claude Shannon, who introduced the following *stochastic entropy*:

Definition 3.1.1 (Shannon's entropy). *Starting with a number of necessary properties termed as* ***Shannon's postulates****, Claude Shannon succeeded to show that there is an essentially unique function H which meets all these desired properties:*

$$H : \mathbb{P} \to \mathbb{R} \tag{3.1}$$

$$H\left(P_{Y|\{d_Y\}}\right) = \sum_{i=1}^{N} P_{Y|\{d_Y\}}(\{y_i\}) \log\left(\frac{1}{P_{Y|\{d_Y\}}(\{y_i\})}\right) \tag{3.2}$$

In **physical systems** the function H had been introduced by Boltzmann [30] as a **measure of disorder**. In **communication systems**, the uncertainty about the actual message to be transmitted, is called **entropy of the source**. Moreover, let us define the differential entropy as follows:

Definition 3.1.2 (Differential entropy). *If $P_{Y|\{d_Y\}}$ is replaced by a continuous density*[1] *with range of variability $\mathcal{X}_Y(\{d_Y\}) = \{y \,|\, a \leq y \leq b\}$ where $a = y_1$ and $b = y_N$ and density function $f_{Y|\{d_Y\}}$, then the stochastic entropy may be represented by*

$$H\left(f_{Y|\{d_Y\}}\right) = \int_a^b f_{Y|\{d_Y\}}(y) \log\left(\frac{1}{f_{Y|\{d_Y\}}(y)}\right) dy \tag{3.3}$$

Nextly, we shall proceed to discuss the Shannon's entropy, the proof of which necessitates the usage of four postulates (referred to [33], pages 547 and 548).

[1]The entropy with regard to a continuous density has many of the properties of the discrete entropy. But, unlike the entropy of a discrete probability distribution, the same of a continuous probability may be positive, infinitely large or even negative [2]. The entropy of a discrete distribution remains invariant with respect to a transformation of a random variable. However, with subject to a continuous random variable, the entropy does not necessarily remain invariant. This is one of the existing difficulties in continuous cases.

3.2 The Shannon's measure of entropy

3.2.1 The Shannon's postulates

Let E_N be the representation of a discrete probability distribution with a finite and bounded support $\mathcal{X}_Y = \{y_1, y_2, \ldots, y_N\}$ given by

$$E_N = (p_1, p_2, \ldots, p_N) \text{ with } p_i = P_{Y|\{d_Y\}}(\{y_i\}) \in [0,1], N \in \mathbb{N} \tag{3.4}$$

then the uncertainty of the probability distribution $P_{Y|\{d_Y\}}$ denoted by $H(E_N)$ must meet the following **four** postulates (referred to [33], pages 547 and 548):

Statement 3.2.1 (Postulate I). *The entropy $H(E_N)$ depends only on the probability distribution denoted by E_N, consequently, it will be denoted by $H(p_1, p_2, \ldots, p_N)$. Additionally, $H(p_1, p_2, \ldots, p_N)$ is a symmetric function of its arguments $p_1, p_2, \ldots, p_N$.*
*(**Note**, that the probability distribution denoted by E_N contains exactly N probability elements.)*

Statement 3.2.2 (Postulate II). *$H(p, 1-p)$ is a continuous function of p $(0 \leq p \leq 1)$.*

Statement 3.2.3 (Postulate III). *$H\left(\frac{1}{2}, \frac{1}{2}\right) = \log_c 2$, such that $c > 1$.*
*(**Note**, that, referred to the page 548 of [33], the value of c has been taken to be 2. But, we shall go for the general case for $c > 1$.)*

Statement 3.2.4 (Postulate IV).

$$\begin{aligned} &H(p_1, p_2, \ldots, p_N) \\ &= H(p_1 + p_2, p_3, p_4, \ldots, p_N) + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \end{aligned}$$

In the year 1948, **Claude E. Shannon** derived the aforesaid function H and showed that its expression is essentially **unique**. He called it **entropy**. This very assertion proposed by Shannon can be formulated in form of a theorem termed as **Shannon's theorem**, namely the **theorem 3.2.2** stated and proved in the **subsection 3.2.3**.

The **proof** of the **Shannon's theorem** necessitates the need to state and prove the **theorem 3.2.1** (this theorem happens to be the stated and proved

theorem 2 given in pages 544 - 546 of [33] [2]). This **theorem 3.2.1** we shall term as the **lemma** for the **Shannon's theorem**, which is discussed as follows:

3.2.2 The lemma for the Shannon's theorem

Theorem 3.2.1 (The lemma for Shannon's theorem). *The function $H(E_N) = \log_2 N$ (i.e. function of the natural number N) is the only function, which satisfies the following postulates (A^*), (B^*) and (C):*

- *(**A***): If the natural numbers N and M are prime to each other, then $H(E_{NM}) = H(E_N) + H(E_M)$*
- *(**B***): $\lim_{N\to\infty} (H(E_{N+1}) - H(E_N)) = 0$.*
- *(**C**): $H(E_2) = 1$*

*Proof of the **theorem 3.2.1**.* At the very first step, we must introduce a trivially taken concept, which shall be unavoidably of use: Hypothetically speaking, $N = 0$ can only be interpreted as the nonexistence of any probability distribution and hence there cannot be any question of any uncertainty of the nonexistent probability distribution. This can be interpreted as $H(E_N) = H(E_0) = 0$.

Let a natural number $S > 1$ be either any prime number or any integral power of any prime number and $f(N) = H(E_N)$ a function (of N) satisfying the postulates (A^*), (B^*) and (C). Obviously, the following must hold:

$$f(0) = 0 \tag{3.5}$$

Let us construct a function $g(N)$ (of N) in the following manner:

$$g(N) = f(N) - \frac{f(S) LOG_2 N}{\log_2 S} \tag{3.6}$$

such that

$$LOG_2 N = \begin{cases} \log_2 N, \text{ if } N > 0 \\ \\ 0, \text{ if } N = 0 \end{cases} \tag{3.7}$$

[2] the proof of this theorem 2 is not presented with enough clarity in the pages 544 - 546 of [33]

Thus, with subject to (3.5) and (3.7), the special case for (3.6) is given as

$$g(0) = 0 \tag{3.8}$$

Clearly, for $N > 0$, the function $g(N)$ fulfills the postulate (A^*), as

$$\begin{aligned} g(NM) &= f(NM) - \frac{f(S)\log_2(NM)}{\log_2 S} \\ &= f(N) + f(M) - \frac{(\log_2 N + \log_2 M)}{\log_2 S} f(S) \\ &= g(N) + g(M) \end{aligned} \tag{3.9}$$

Now, for $N > 0$, let us put

$$\epsilon_N = g(N+1) - g(N) = f(N+1) - f(N) - \frac{f(S)}{\log_2 S}\log_2\left(1 + \frac{1}{N}\right) \tag{3.10}$$

Clearly, for $N > 0$, $g(N)$ fulfills B^*, simply because $f(N)$ fulfills B^*, which is presented as

$$\lim_{N\to\infty} \epsilon_N = \lim_{N\to\infty} (f(N+1) - f(N)) - \frac{f(S)}{\log_2 S}\log_2(1+0) = 0 \tag{3.11}$$

Here, we see that

$$g(S) = f(S) - \frac{f(S)\log_2 S}{\log_2 S} = 0 \tag{3.12}$$

Now, for any $N > 0$, let us define a natural number $N^{(1)}$ by

$$N^{(1)} = \begin{cases} \left[\frac{N}{S}\right], & \text{if } GCF\left(\left[\frac{N}{S}\right], S\right) = 1 \\ \left[\frac{N}{S}\right] - 1, & \text{if } GCF\left(\left[\frac{N}{S}\right], S\right) > 1 \end{cases} \tag{3.13}$$

such that $GCF(k_1, k_2)$ denotes the **g**reatest **c**ommon **f**actor of the two natural numbers k_1 and k_2.

Clearly, $N^{(1)} \leq \frac{N}{S} < N^{(1)} + 2$ and therefore let us define $\ell^{(1)} = N - N^{(1)}S$.

Therefore, $N^{(1)}S + 2S > N = N^{(1)}S + \ell^{(1)}$ gives

$$2S > \ell^{(1)} \geq 0 \tag{3.14}$$

Thus, by (3.9),

$$\begin{aligned} g(N^{(1)}S) &= g(S) + g(N^{(1)}) \\ &= 0 + g(N^{(1)}) \qquad \text{(by 3.12)} \\ &= g(N^{(1)}) \end{aligned} \tag{3.15}$$

which leads us to

$$\begin{aligned} g(N) &= g(N) + g(N^{(1)}) - g(N^{(1)}S) \\ &= g(N^{(1)}) + \big(g(N) - g(N^{(1)}S)\big) \\ &= g(N^{(1)}) + [\{g(N) - g(N-1)\} + \{g(N-1) - g(N-2)\} \\ &\qquad + \ldots + \big\{g(N^{(1)}S+1) - g(N^{(1)}S)\big\}] \\ &= g(N^{(1)}) + \underbrace{\sum_{k=N^{(1)}S}^{N-1} \epsilon_k}_{\ell^{(1)} \text{ terms } < 2S \text{ terms (by 3.14)}} \end{aligned} \tag{3.16}$$

Precisely, the finite sum contained in (3.16), namely $\sum_{k=N^{(1)}S}^{N-1} \epsilon_k$ contains $\ell^{(1)}$ terms with $\ell^{(1)} < 2S$.

Again, by defining the natural number $N^{(2)}$ exactly in the similar manner as in the case of $N^{(1)}$, viz

$$N^{(2)} = \begin{cases} \left[\frac{N^{(1)}}{S}\right], & \text{if } GCF\left(\left[\frac{N^{(1)}}{S}\right], S\right) = 1 \\ \left[\frac{N^{(1)}}{S}\right] - 1, & \text{if } GCF\left(\left[\frac{N^{(1)}}{S}\right], S\right) > 1 \end{cases} \tag{3.17}$$

so that $N^{(2)} \leq \frac{N^{(1)}}{S} < N^{(2)} + 2$ and consequently $N^{(2)} \leq \frac{N^{(1)}}{S} \leq \frac{N}{S^2}$ (because of $N^{(1)} < \frac{N}{S}$) and let us similarly define $\ell^{(2)} = N^{(1)} - N^{(2)}S$.

In the same way, by $\frac{N^{(1)}}{S} < N^{(2)} + 2$, $N^{(2)}S + 2S > N^{(1)} = N^{(2)}S + \ell^{(2)}$ gives

$$2S > \ell^{(2)} \geq 0 \tag{3.18}$$

Exactly in the same way we derived the relation between $g(N)$ and $g(N^{(1)})$,

namely (3.16), we get the relation between $g(N^{(1)})$ and $g(N^{(2)})$ as

$$g(N^{(1)}) = g(N^{(2)}) + \underbrace{\sum_{k=N^{(2)}S}^{N^{(1)}-1} \epsilon_k}_{\ell^{(2)} \text{ terms} < 2S \text{ terms (by 3.18)}} \tag{3.19}$$

Precisely, the finite sum contained in (3.19), namely $\sum_{k=N^{(2)}S}^{N^{(1)}-1} \epsilon_k$ contains $\ell^{(2)}$ terms with $\ell^{(2)} < 2S$.

Therefore, by applying the expression of $g(N^{(1)})$ in (3.19) on (3.16), we get

$$g(N) = g(N^{(2)}) + \underbrace{\sum_{k=N^{(1)}S}^{N-1} \epsilon_k}_{<2S \text{ terms}} + \underbrace{\sum_{k=N^{(2)}S}^{N^{(1)}-1} \epsilon_k}_{<2S \text{ terms}} \tag{3.20}$$

with subject to $N^{(2)} \leq \frac{N^{(1)}}{S} \leq \frac{N}{S^2}$.

Proceeding exactly in the same way, we get in the very next step

$$g(N) = g(N^{(3)}) + \underbrace{\sum_{k=N^{(1)}S}^{N-1} \epsilon_k}_{<2S \text{ terms}} + \underbrace{\sum_{k=N^{(2)}S}^{N^{(1)}-1} \epsilon_k}_{<2S \text{ terms}} + \underbrace{\sum_{k=N^{(3)}S}^{N^{(2)}-1} \epsilon_k}_{<2S \text{ terms}} \tag{3.21}$$

with subject to $N^{(3)} \leq \frac{N^{(2)}}{S} \leq \frac{N}{S^3}$.

Proceeding exactly in the same way, we get in the j^{th} step

$$g(N) = g(N^{(j)}) + \underbrace{\sum_{k=N^{(1)}S}^{N-1} \epsilon_k}_{<2S \text{ terms}} + \underbrace{\sum_{k=N^{(2)}S}^{N^{(1)}-1} \epsilon_k}_{<2S \text{ terms}} + \ldots + \underbrace{\sum_{k=N^{(j)}S}^{N^{(j-1)}-1} \epsilon_k}_{<2S \text{ terms}} \tag{3.22}$$

with subject to $N^{(j)} \leq \frac{N}{S^j}$.

Now, at this point, we must have a close look at the following: For every natural number N (obviously $N > 0$), there exists a natural number j, such

that $\frac{N}{S^j} < 1$. In that case, $N^{(j)} = 0$ and thereby with the reference to (3.8), we have

$$g(N^{(j)}) = 0 \tag{3.23}$$

At the same time, $\frac{N}{S^j} < 1 \Leftrightarrow \log_S N < j$ necessarily means that $N^{(j)} = 0$ is reached after at most $[\log_S N] + 1$ procedural steps described immediately above, i.e. when j reaches the value $[\log_S N] + 1$.

(3.24)

Therefore, by applying (3.23) on (3.22), we get

$$g(N) = \underbrace{\sum_{k=N^{(1)}S}^{N-1} \epsilon_k}_{<2S \text{ terms}} + \underbrace{\sum_{k=N^{(2)}S}^{N^{(1)}-1} \epsilon_k}_{<2S \text{ terms}} + \ldots + \underbrace{\sum_{k=N^{(j)}S}^{N^{(j-1)}-1} \epsilon_k}_{<2S \text{ terms}} \tag{3.25}$$

and with subject to (3.24), the expression of $g(N)$, viz. (3.25) can contain at most (in fact **strictly** less than) $2S([\log_S N] + 1)$ terms.

Moreover, if $\epsilon_{N,\max}$ and $\epsilon_{N,\min}$ are the maximum and the minimum values among all the above ϵ values present in (3.25) respectively, then three cases do arise and in each of these three cases, by (3.11) we shall make use of $\lim_{N\to\infty} \epsilon_{N,\max} = \lim_{N\to\infty} \epsilon_{N,\min} = 0$:

- **Case 1**: ($\epsilon_{N,\min} \geq 0$ and $\epsilon_{N,\max} \geq 0$):

  Here, $0 \leq g(N) \leq 2S([\log_S N] + 1)\epsilon_{N,\max}$

- **Case 2**: ($\epsilon_{N,\min} \leq 0$ and $\epsilon_{N,\max} \geq 0$):

  Here, $2S([\log_S N] + 1)\epsilon_{N,\min} \leq g(N) \leq 2S([\log_S N] + 1)\epsilon_{N,\max}$

- **Case 3**: ($\epsilon_{N,\min} \leq 0$ and $\epsilon_{N,\max} \leq 0$):

  Here, $2S([\log_S N] + 1)\epsilon_{N,\min} \leq g(N) \leq 0$

Thus, by considering each of the above three cases, we can easily arrive at

$$\lim_{N\to\infty} \frac{g(N)}{\log_2 N} = 0 \tag{3.26}$$

and hence by (3.6), we get

$$\lim_{N\to\infty} \frac{f(N)}{\log_2 N} = \frac{f(S)}{\log_2 S} \tag{3.27}$$

which necessarily means that the constant $\frac{f(S)}{\log_2 S}$ is independent of S and therefore let us take $\frac{f(S)}{\log_2 S} = c_0$.

Thus,

$$f(S) = c_0 \log_2 S \tag{3.28}$$

Now, for any natural number N with $N > 1$, which is a composition in form of a product of $S_1, S_2, \ldots, S_r$, namely $N = S_1 S_2 \ldots S_r$, such that each S_i, $i = 1, 2, \ldots, r$ is either a prime number or an integral power of a prime number, we have

$$\begin{aligned} f(N) &= f(S_1 S_2 \ldots S_r) \\ &= \sum_{i=1}^{r} f(S_i) \ \ (\text{ since } f(S_i) = H(E_{S_i}) \text{ satisfies the postulate } (A^*) \ ) \\ &= \sum_{i=1}^{r} c_0 \log_2 S_i \ \ (\text{ by } (3.28) \ ) \\ &= c_0 \log_2 N \end{aligned} \tag{3.29}$$

Again, by postulate (C^*), $f(2) = H(E_2) = 1$ and thus by (3.29), we have $f(2) = 1 = c_0 \log_2 2$, which brings us to $c_0 = 1$.

Hence, (3.29) is rewritten as

$$f(N) = \log_2 N \tag{3.30}$$

and this completes the proof of the **theorem 3.2.1**. □

3.2.3 The Shannon's theorem

Theorem 3.2.2 (Shannon's theorem). *:*
With subject to the fulfillment of the postulates (I), (II), (III) and (IV) as stated in the ***subsection 3.2.1****, H is uniquely given by*

$$H(E_N) = \sum_{i=1}^{N} p_i \log_c \left(\frac{1}{p_i}\right) \text{ for } p_i = P_{Y|\{d_Y\}}(\{y_i\}) \in [0,1], N \in \mathbb{N} \quad (3.31)$$

such that the logarithmic base, namely c, is any fixed real value strictly greater than 1.

In other words, the expression of H given by (3.31) is the ***only*** *expression, which fulfills the postulates (I), (II), (III) and (IV).*

Proof of the ***theorem 3.2.2****.* The formal derivation of the Shannon's Entropy (3.31) necessitates the usage of postulates (I), (II), (III) and (IV) as well as the usage of the **theorem 3.2.1** (belonging to the **subsection 3.2.2**). This formal derivation shall be given in **six** broad steps:

Step 1:

At first, we shall show that $H(1) = H(E_1) = 0$, i.e. the entropy of a single point probability distribution is zero.

For $N = 2$ and together with $p_1 = 1, p_2 = 0$, by the postulate (IV), we have

$$H(1,0) = H(1+0) + (1+0)H\left(\frac{1}{1+0}, \frac{0}{1+0}\right) = H(1) + H(1,0)$$

which gives

$$H(1) = 0 \quad (3.32)$$

Then, we shall show that the zero probability elements do not change the entropy of the probability distribution.

By postulate (I) subjecting to the symmetric property of the entropy, we have

$$H(p_1, p_2, \ldots, p_N, 0) = H(0, p_1, p_2, \ldots, p_N) \quad (3.33)$$

and by postulate (IV),

$$\begin{aligned} H(0, p_1, p_2, \ldots, p_N) &= H(0 + p_1, p_2, p_3, \ldots, p_N) + (0 + p_1) H\left(\frac{0}{0 + p_1}, \frac{p_1}{0 + p_1}\right) \\ &= H(p_1, p_2, \ldots, p_N) + p_1 H(0, 1) \end{aligned} \tag{3.34}$$

Now, let us make a careful observation of the following: Basically, the degenerated probability distribution defined by $E_1 : \{p_1 = 1\}$ is completely identical with the degenerated probability distribution defined by $E_2 : \{p_1 = 0, p_2 = 1\}$. This necessarily means that the corresponding entropies $H(E_1) = H(1)$ and $H(E_2) = H(0, 1)$ must be identically the same, i.e. $H(1) = H(0, 1)$ and therefore by (3.32), we conclude

$$H(0, 1) = 0 \tag{3.35}$$

and therefore (3.34) can be rewritten with subject to (3.35) as

$$H(0, p_1, p_2, \ldots, p_N) = H(p_1, p_2, \ldots, p_N) \tag{3.36}$$

Of course, by (3.36) and (3.33), we can also write

$$H(p_1, p_2, \ldots, p_N, 0) = H(p_1, p_2, \ldots, p_N) \tag{3.37}$$

which proves that the **zero probability elements do not contribute to the Shannon's entropy**.

Step 2:

In this step, we shall generalize the recursiveness stated in the postulate (IV) by proving the following relation:

$$\begin{aligned} &H(p_1, p_2, \ldots, p_{N_1}, p_{N_1+1}, \ldots, p_{N_1+N_2}) \\ &= H(s_{N_1}, p_{N_1+1}, p_{N_1+2}, \ldots, p_{N_1+N_2}) + s_{N_1} H\left(\frac{p_1}{s_{N_1}}, \frac{p_2}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) \\ &\text{such that } s_{N_1} = \sum_{j=1}^{N_1} p_j \end{aligned} \tag{3.38}$$

We shall give the proof of (3.38) by the mathematical induction with respect to the natural number N_1.

Clearly, the relation (3.38) is fulfilled for $N_1 = 2$, which can be easily shown by rewriting the very postulate (IV) for $2 + N_2$ probability elements, so that $s_2 = p_1 + p_2$, in the following way

$$H(p_1, p_2, \ldots, p_{2+N_2}) = H(s_2, p_{2+1}, \ldots, p_{2+N_2}) + s_2 H\left(\frac{p_1}{s_2}, \frac{p_2}{s_2}\right) \tag{3.39}$$

Now, for the purpose of the induction step, we shall assume that the relation (3.38) is valid for $N_1 - 1$ at the place of N_1, after which shall be in a position to prove that (3.38) holds for N_1. Here, the said validity holds for the $N_1 + N_2 - 1$ probability elements $p_1 + p_2, p_3, p_4, \ldots, p_{N_1+N_2}$ can be expressed as

$$\begin{aligned} &H(p_1 + p_2, p_3, \ldots, p_{N_1}, p_{N_1+1}, \ldots, p_{N_1+N_2}) \\ &= H\left(\underbrace{(p_1 + p_2) + p_3 + \ldots + p_{N_1}}_{=s_{N_1}}, p_{N_1+1}, \ldots, p_{N_1+N_2}\right) \\ &\quad + s_{N_1} H\left(\frac{p_1 + p_2}{s_{N_1}}, \frac{p_3}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) \end{aligned} \tag{3.40}$$

Now, by using the postulate (IV), we have the following two relations:

$$\begin{aligned} &H(p_1, p_2, \ldots, p_{N_1+N_2}) \\ &= H((p_1 + p_2), p_3, \ldots, p_{N_1+N_2}) + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \end{aligned} \tag{3.41}$$

and

$$\begin{aligned} &H\left(\frac{p_1}{s_{N_1}}, \frac{p_2}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) \\ &= H\left(\frac{p_1 + p_2}{s_{N_1}}, \frac{p_3}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) + \left(\frac{p_1}{s_{N_1}} + \frac{p_2}{s_{N_1}}\right) H\left(\frac{\frac{p_1}{s_{N_1}}}{\frac{p_1}{s_{N_1}} + \frac{p_2}{s_{N_1}}}, \frac{\frac{p_2}{s_{N_1}}}{\frac{p_1}{s_{N_1}} + \frac{p_2}{s_{N_1}}}\right) \\ &= H\left(\frac{p_1 + p_2}{s_{N_1}}, \frac{p_3}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) + \left(\frac{p_1 + p_2}{s_{N_1}}\right) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \end{aligned} \tag{3.42}$$

Now, by replacing the expression $H\left((p_1+p_2), p_3, \ldots, p_{N_1+N_2}\right)$ existing in (3.41) by the same given by (3.40), we rewrite (3.41) as

$$\begin{aligned} & H(p_1, p_2, \ldots, p_{N_1+N_2}) \\ & = H(s_{N_1}, p_{N_1+1}, \ldots, p_{N_1+N_2}) + s_{N_1} H\left(\frac{p_1+p_2}{s_{N_1}}, \frac{p_3}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) \\ & \quad + (p_1+p_2) H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right) \end{aligned} \tag{3.43}$$

Now, for the sake of convenience, let us rewrite (3.42) in the following way:

$$\begin{aligned} & s_{N_1} H\left(\frac{p_1+p_2}{s_{N_1}}, \frac{p_3}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) \\ & = s_{N_1} H\left(\frac{p_1}{s_{N_1}}, \frac{p_2}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) - (p_1+p_2) H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right) \end{aligned} \tag{3.44}$$

and hence, by replacing the expression of $s_{N_1} H\left(\frac{p_1+p_2}{s_{N_1}}, \frac{p_3}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right)$ existing in (3.43) by the same given by (3.44), we finally rewrite (3.43) as

$$\begin{aligned} & H(p_1, p_2, \ldots, p_{N_1+N_2}) \\ & = H(s_{N_1}, p_{N_1+1}, \ldots, p_{N_1+N_2}) \\ & \quad + s_{N_1} H\left(\frac{p_1}{s_{N_1}}, \frac{p_2}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) - (p_1+p_2) H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right) \\ & \quad + (p_1+p_2) H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right) \\ & = H(s_{N_1}, p_{N_1+1}, \ldots, p_{N_1+N_2}) + s_{N_1} H\left(\frac{p_1}{s_{N_1}}, \frac{p_2}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) \end{aligned} \tag{3.45}$$

which is precisely our relation (3.38). In other words, by mathematical induction, our assertion (3.38) is proved.

Step 3:

In this step, we shall generalize the recursiveness of postulate (IV) still further by proving the following relation:

$$\begin{aligned} & H(p_1^{(1)}, \ldots, p_{N_1}^{(1)}, p_1^{(2)}, \ldots, p_{N_2}^{(2)}, \ldots, p_1^{(M)}, \ldots, p_{N_M}^{(M)}) \\ & = H(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M) + \sum_{j=1}^{M} \mathbf{s}_j H\left(\frac{p_1^{(j)}}{\mathbf{s}_j}, \frac{p_2^{(j)}}{\mathbf{s}_j}, \ldots, \frac{p_{N_j}^{(j)}}{\mathbf{s}_j}\right) \\ & \text{such that } \mathbf{s}_j = \sum_{k=1}^{N_j} p_k^{(j)} \text{ together with } \sum_{j=1}^{M} \mathbf{s}_j = 1 \end{aligned} \tag{3.46}$$

For this, by using (3.38), the expression on the left hand side of (3.46) can be written as (*purely for the sake of convenience we shall proceed from the next page*):

$$
\begin{aligned}
& H(p_1^{(1)}, \ldots, p_{N_1}^{(1)}, p_1^{(2)}, \ldots, p_{N_2}^{(2)}, \ldots, p_1^{(M)}, \ldots, p_{N_M}^{(M)}) \\
& = H(\mathbf{s}_1, p_1^{(2)}, \ldots, p_{N_2}^{(2)}, \ldots, p_1^{(M)}, \ldots, p_{N_M}^{(M)}) \\
& \quad + \mathbf{s}_1 H\left(\frac{p_1^{(1)}}{\mathbf{s}_1}, \frac{p_2^{(1)}}{\mathbf{s}_1}, \ldots, \frac{p_{N_1}^{(1)}}{\mathbf{s}_1}\right) \\
& = H(p_1^{(2)}, \ldots, p_{N_2}^{(2)}, \ldots, p_1^{(M)}, \ldots, p_{N_M}^{(M)}, \mathbf{s}_1) \\
& \quad + \mathbf{s}_1 H\left(\frac{p_1^{(1)}}{\mathbf{s}_1}, \frac{p_2^{(1)}}{\mathbf{s}_1}, \ldots, \frac{p_{N_1}^{(1)}}{\mathbf{s}_1}\right) \\
& \quad \text{(by postulate } I \text{ with regard to the symmetry)} \\
& = H(\mathbf{s}_2, p_1^{(3)}, \ldots, p_{N_2}^{(3)}, \ldots, p_1^{(M)}, \ldots, p_{N_M}^{(M)}, \mathbf{s}_1) \\
& \quad + \mathbf{s}_2 H\left(\frac{p_1^{(2)}}{\mathbf{s}_2}, \frac{p_2^{(2)}}{\mathbf{s}_2}, \ldots, \frac{p_{N_2}^{(2)}}{\mathbf{s}_2}\right) + \mathbf{s}_1 H\left(\frac{p_1^{(1)}}{\mathbf{s}_1}, \frac{p_2^{(1)}}{\mathbf{s}_1}, \ldots, \frac{p_{N_1}^{(1)}}{\mathbf{s}_1}\right) \\
& \quad \text{(by using (3.38) again)} \\
& = H(p_1^{(3)}, \ldots, p_{N_2}^{(3)}, \ldots, p_1^{(M)}, \ldots, p_{N_M}^{(M)}, \mathbf{s}_1, \mathbf{s}_2) \\
& \quad + \mathbf{s}_1 H\left(\frac{p_1^{(1)}}{\mathbf{s}_1}, \frac{p_2^{(1)}}{\mathbf{s}_1}, \ldots, \frac{p_{N_1}^{(1)}}{\mathbf{s}_1}\right) + \mathbf{s}_2 H\left(\frac{p_1^{(2)}}{\mathbf{s}_2}, \frac{p_2^{(2)}}{\mathbf{s}_2}, \ldots, \frac{p_{N_2}^{(2)}}{\mathbf{s}_2}\right) \\
& \quad \text{(by postulate } I \text{ with regard to the symmetry once again)}
\end{aligned}
\tag{3.47}
$$

Proceeding exactly in this way, we shall get the following expression in the final step of (3.47) as

$$
H(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M) + \sum_{j=1}^{M} \mathbf{s}_j H\left(\frac{p_1^{(j)}}{\mathbf{s}_j}, \frac{p_2^{(j)}}{\mathbf{s}_j}, \ldots, \frac{p_{N_j}^{(j)}}{\mathbf{s}_j}\right)
$$

and this is precisely the proof of our relation (3.46).

Step 4:

Here, if U_N represents a discrete uniform distribution with N probability elements, then by setting

$$f(N) = H(U_N) = H\underbrace{\left(\frac{1}{N}, \ldots, \frac{1}{N}\right)}_{N \text{ elements}} \tag{3.48}$$

our objective shall be to show that

$$f(NM) = f(N) + f(M), \text{ for all } N, M \in \mathbb{N} \tag{3.49}$$

In order to do this, we shall use the derived relation (3.46) by putting $N_j = N$ for every $j \in \{1, 2, \ldots, M\}$. In that case,

- $p_k^{(j)} = \frac{1}{NM}$ for every $k \in \{1, 2, \ldots, N\}$
- $\mathbf{s}_j = \mathbf{s} = \sum_{k=1}^{N} p_k^{(j)} = N\frac{1}{NM} = \frac{1}{M}$

and therefore by (3.46),

$$\begin{aligned} H\underbrace{\left(\frac{1}{NM}, \ldots, \frac{1}{NM}\right)}_{NM \text{ elements}} &= H\underbrace{(\mathbf{s}, \ldots, \mathbf{s})}_{M \text{ elements}} + \sum_{j=1}^{M} \mathbf{s} H\underbrace{\left(\frac{\frac{1}{NM}}{\mathbf{s}}, \ldots, \frac{\frac{1}{NM}}{\mathbf{s}}\right)}_{N \text{ elements}} \\ &= H\underbrace{\left(\frac{1}{M}, \ldots, \frac{1}{M}\right)}_{M \text{ elements}} + M\,\frac{1}{M}\,H\underbrace{\left(\frac{1}{N}, \ldots, \frac{1}{N}\right)}_{N \text{ elements}} \\ &= H\underbrace{\left(\frac{1}{M}, \ldots, \frac{1}{M}\right)}_{M \text{ elements}} + H\underbrace{\left(\frac{1}{N}, \ldots, \frac{1}{N}\right)}_{N \text{ elements}} \\ &= f(M) + f(N) \end{aligned} \tag{3.50}$$

which proves nothing, but the very asserted relation (3.49).

Step 5:

In this step, we shall show at first the following recursive relation

$$f(N) = H\left(\frac{1}{N}, 1-\frac{1}{N}\right) + \left(1-\frac{1}{N}\right) f(N-1) \tag{3.51}$$

where the definition of $f(N)$ is subjected to (3.48). Only after that, by setting

$$d_N = f(N) - f(N-1) \tag{3.52}$$

we shall proceed to show that

$$\lim_{N\to\infty} d_N = 0 \tag{3.53}$$

Now, by setting

- $N_1 = N-1$, $N_2 = N$
- $p_j = \frac{1}{N}$ for every $j \in \{1, 2, \ldots, N_2\}$
- and therefore, $s_{N_1} = \sum_{j=1}^{N_1} \frac{1}{N} = 1 - \frac{1}{N}$ and
- $f(N) = H\left(p_1, p_2, \ldots, p_{N_1}, p_{N_2}\right)$

in the relation (3.38), we get

$$\begin{aligned}
&H\left(p_1, p_2, \ldots, p_{N_1}, p_{N_2}\right) \\
&= H(s_{N_1}, p_{N_2}) + s_{N_1} H\left(\frac{p_1}{s_{N_1}}, \frac{p_1}{s_{N_1}}, \ldots, \frac{p_{N_1}}{s_{N_1}}\right) \\
&= H\left(1-\frac{1}{N}, \frac{1}{N}\right) + \left(1-\frac{1}{N}\right) H\underbrace{\left(\frac{\frac{1}{N}}{1-\frac{1}{N}}, \ldots, \frac{\frac{1}{N}}{1-\frac{1}{N}}\right)}_{N-1 \text{ elements}} \qquad (3.54) \\
&= H\left(\frac{1}{N}, 1-\frac{1}{N}\right) + \left(1-\frac{1}{N}\right) f(N-1) \\
&\quad \text{(by postulate } I \text{ with regard to the symmetry)}
\end{aligned}$$

and thus our asserted relation (3.51) is proved.

Now, we proceed to prove (3.53). At first, we set

$$\delta_N = H\left(\frac{1}{N}, 1-\frac{1}{N}\right) \tag{3.55}$$

which gives

- $\lim_{N\to\infty} \delta_N = H(0,1) = 0$ by the postulate (II) saying about the continuity of $H(p, 1-p)$ for $0 \leq p \leq 1$ as well as by (3.35)
- $\delta_1 = H(1,0) = H(0,1) = 0$ by postulate (I) with regard to the symmetry as well as by (3.35)
- by postulate (II), δ_N is continuous for every $N \in \mathbb{N} \setminus \{0\}$

and this can only mean that $|\delta_N|$ is purely a function of N and is bounded above. Thus, $|\delta_N|$ has to have a global maximum value within the range of $1 \leq N < \infty$.

So, let us take

$$\delta_{\max} = \max_{N\in\mathbb{N}} |\delta_N| \tag{3.56}$$

Notably, $\delta_{\max} = 0$ would trivially mean that $H\left(\frac{1}{N}, 1-\frac{1}{N}\right) = 0$ for every $n \in \mathbb{N}$, which is absurd. Thus, we will have to conclude that $\delta_{\max} > 0$.

Now, by the definition (3.52) of d_N,

$$\begin{aligned} & d_2 + d_3 + \ldots + d_{N-1} \\ & = (f(2) - f(1)) + (f(3) - f(2)) + \ldots + (f(N-1) - f(N-2)) \\ & = f(N-1) - f(1) \\ & = f(N-1) - H(1) \\ & = f(N-1) \text{ (by (3.32))} \end{aligned} \tag{3.57}$$

which is rewritten as

$$f(N-1) = d_2 + d_3 + \ldots + d_{N-1} \tag{3.58}$$

Again, by using the definition (3.55) of δ_N, the recursive relation (3.51) can

be rewritten as

$$\begin{aligned}\delta_N &= f(N) - \left(1 - \frac{1}{N}\right) f(N-1) \\ &= d_N + \frac{1}{N} f(N-1) \quad (\text{ by (3.52) }) \\ &= d_N + \frac{1}{N}(d_2 + d_3 + \ldots + d_{N-1}) \quad (\text{ by (3.58) })\end{aligned} \tag{3.59}$$

which is again conveniently rewritten as

$$N\delta_N = Nd_N + d_2 + d_3 + \ldots + d_{N-1} \tag{3.60}$$

Before we go ahead, we make a note of the following: With regard to (3.5) and (3.32), we have $f(0) = 0$ and $f(1) = H(1) = 0$ and this brings us to

$$d_1 = f(1) - f(0) = 0 \tag{3.61}$$

So, with regard to (3.60) and (3.61), we calculate

$$\begin{aligned}\sum_{j=2}^{N} j\delta_j &= \sum_{j=2}^{N}(jd_j + d_1 + d_2 + d_3 + \ldots + d_{j-1}) \\ &= (2d_2 + d_1) + (3d_3 + d_2) + (4d_4 + d_2 + d_3) \\ &\quad + \ldots + (Nd_N + d_2 + d_3 + \ldots + d_{N-1}) \\ &= (2d_2 + (N-2)d_2) + (3d_3 + (N-3)d_3) + (4d_4 + (N-4)d_4) \\ &\quad + \ldots + (Nd_N + (N-N)d_N) \\ &= N(d_2 + d_3 + \ldots + d_N)\end{aligned} \tag{3.62}$$

which in turn is equivalently and conveniently can be rewritten as

$$\frac{2}{N+1}\sum_{k=2}^{N} d_k = \frac{\sum_{j=2}^{N} j\delta_j}{\sum_{j=1}^{N} j} \tag{3.63}$$

Our immediately next step will be to show that the limiting value of right hand side of (3.63) for $N \to \infty$ is equal to zero, i.e.

$$\lim_{N\to\infty} \frac{\sum_{j=2}^{N} j\delta_j}{\sum_{j=1}^{N} j} = 0 \tag{3.64}$$

For this, because of $\lim_{N\to\infty} \delta_N = H(0,1) = 0$ (by (3.35)), we conclude:

For an arbitrarily small $\epsilon > 0$, there exists a natural number N_ϵ, such that $0 \leq |\delta_N| < \frac{\epsilon}{2}$ for every $N > N_\epsilon$ and therefore,

$$\begin{aligned}
\frac{\sum_{j=2}^{N} j\,|\delta_j|}{\sum_{j=1}^{N} j} &\leq \frac{\sum_{j=2}^{N_\epsilon} j\,|\delta_j| + \frac{\epsilon}{2}\sum_{j=N_\epsilon+1}^{N} j}{\frac{N(N+1)}{2}} \\
&\overset{(3.56)}{\leq} \frac{\delta_{\max}\ \frac{1}{2}N_\epsilon(N_\epsilon+1) + \frac{\epsilon}{2}\left(\frac{1}{2}N(N+1) - \frac{1}{2}N_\epsilon(N_\epsilon+1)\right)}{\frac{N(N+1)}{2}} \\
&= \delta_{\max}\frac{N_\epsilon(N_\epsilon+1)}{N(N+1)} + \frac{\epsilon}{2}\underbrace{\left(1 - \frac{N_\epsilon(N_\epsilon+1)}{N(N+1)}\right)}_{<1 \text{ for } N>N_\epsilon} \\
&< \delta_{\max}\frac{N_\epsilon(N_\epsilon+1)}{N(N+1)} + \frac{\epsilon}{2}
\end{aligned} \tag{3.65}$$

Again, as $\delta_{\max}$ is a fixed positive number, there exists a natural number N'_ϵ such that $\frac{N_\epsilon(N_\epsilon+1)}{N(N+1)} < \frac{\epsilon}{2\delta_{\max}}$ for every $N > N'_\epsilon$ and thus we proceed with the inequality (3.65) as

$$\begin{aligned}
\frac{\sum_{j=2}^{N} j\,|\delta_j|}{\sum_{j=1}^{N} j} &< \delta_{\max}\frac{N_\epsilon(N_\epsilon+1)}{N(N+1)} + \frac{\epsilon}{2} \\
&< \delta_{\max}\frac{\epsilon}{2\delta_{\max}} + \frac{\epsilon}{2} \\
&= \epsilon
\end{aligned} \tag{3.66}$$

Hence, for an arbitrarily chosen ϵ, we have found that $\left|\frac{\sum_{j=2}^{N} j\delta_j}{\sum_{j=1}^{N} j}\right| \leq \frac{\sum_{j=2}^{N} j|\delta_j|}{\sum_{j=1}^{N} j} < \epsilon$ for every $N > N_{\epsilon}^{'}$. This proves limit stated by our assertion (3.64).

Therefore, by applying the assertion (3.64) on (3.63), we get

$$\lim_{N\to\infty} \frac{2}{N+1} \sum_{k=2}^{N} d_k = 0 \tag{3.67}$$

and this helps us to conclude the following by rewriting (3.60) at first and then by using $\lim_{N\to\infty} \delta_N = 0$ subsequently:

$$\begin{aligned}
&\frac{2}{1+\frac{1}{N}}(\delta_N - d_N) = \frac{2}{1+\frac{1}{N}} \frac{\sum_{k=2}^{N} d_k}{N} \\
&\Rightarrow \lim_{N\to\infty} \frac{2}{1+\frac{1}{N}}(\delta_N - d_N) = \lim_{N\to\infty} \frac{2}{N+1} \sum_{k=2}^{N} d_k \\
&\Rightarrow 2 \lim_{N\to\infty} \delta_N - 2 \lim_{N\to\infty} d_N = 0 \\
&\Rightarrow 0 - 2 \lim_{N\to\infty} d_N = 0 \Leftrightarrow \lim_{N\to\infty} d_N = 0
\end{aligned} \tag{3.68}$$

which ultimately proves our assertion (3.53).

At this very point, we observe that the function $f(N)$

- fulfills the postulate (A^*) because of (3.49) belonging to the **step 4**
- fulfills the postulate (B^*) because of (3.53) (i.e. (3.68)) proven immediately above
- fulfills the postulate (C), which can be shown as follows:

  By setting $N = 2$ in the definition of $f(N)$ given in (3.48), we get $f(2) = H\left(\frac{1}{2}, \frac{1}{2}\right)$ and by postulate (III) by setting $c = 2$, we get $H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$, thereby giving $f(2) = H(U_2) = 1$

and hence, **with regard to the proven theorem 3.2.1**, we arrive at

$$f(N) = \log_2 N \tag{3.69}$$

Step 6:

Let us consider the expression $H\left(\frac{N_a}{N_b}, 1-\frac{N_a}{N_b}\right)$, such that N_a and N_b are natural numbers with $N_a < N_b$.

Now, by considering the derived relation (3.46) in the step 3 with subject to the following setting

- $M = 2$
- $N_1 = N_a$
- $N_2 = N_b - N_a$ (i.e. $= N_M$)
- $p_1^{(1)} = p_2^{(1)} = \ldots = p_{N_a}^{(1)} = p_1^{(2)} = p_2^{(2)} = \ldots = p_{N_b-N_a}^{(2)} = \frac{1}{N_b}$
- $\mathbf{s}_1 = \frac{N_a}{N_b}$ and $\mathbf{s}_2 = \frac{N_b-N_a}{N_b}$

and thus, by using (3.69) on (3.46) we get

$$\begin{aligned}
\log_2 N_b &= H\left(p_1^{(1)}, p_2^{(1)}, \ldots, p_{N_a}^{(1)}, p_1^{(2)}, p_2^{(2)}, \ldots, p_{N_b-N_a}^{(2)}\right) \\
&= H\left(\mathbf{s}_1, \mathbf{s}_2\right) + \mathbf{s}_1 H\underbrace{\left(\frac{\frac{1}{N_b}}{\mathbf{s}_1}, \ldots, \frac{\frac{1}{N_b}}{\mathbf{s}_1}\right)}_{N_a \text{ elements}} + \mathbf{s}_2 H\underbrace{\left(\frac{\frac{1}{N_b}}{\mathbf{s}_2}, \ldots, \frac{\frac{1}{N_b}}{\mathbf{s}_2}\right)}_{N_b-N_a \text{ elements}} \\
&= H\left(\frac{N_a}{N_b}, 1-\frac{N_a}{N_b}\right) \\
&\qquad + \frac{N_a}{N_b}\log_2 N_a + \frac{N_b-N_a}{N_b}\log_2(N_b-N_a) \\
&\qquad \text{(by using (3.69) once again)}
\end{aligned} \tag{3.70}$$

which can be easily conveniently rewritten as

$$H\left(\frac{N_a}{N_b}, 1-\frac{N_a}{N_b}\right) = \frac{N_a}{N_b}\log_2 N_b - \frac{N_a}{N_b}\log_2 N_a + \left(1-\frac{N_a}{N_b}\right)\log_2 N_b - \left(1-\frac{N_a}{N_b}\right)\log_2(N_b - N_a)$$

$$= -\frac{N_a}{N_b}\log_2\left(\frac{N_a}{N_b}\right) - \left(1-\frac{N_a}{N_b}\right)\log_2\left(1-\frac{N_a}{N_b}\right) \tag{3.71}$$

Now, by postulate (II), since $H(p, 1-p)$ is a continuous function of p for every real value of $p \in [0,1]$, we can be allowed to extend the relation (3.71) for any real valued p, which means

$$H(p, 1-p) = -p\log_2 p - (1-p)\log_2(1-p) \tag{3.72}$$

At this very point, we can well see that (3.72) shows that the expression of the Shannon's entropy given by (3.31), namely

$$H(p_1, p_2, \ldots, p_N) = \sum_{i=1}^{N} p_i \log_c\left(\frac{1}{p_i}\right) \text{ for } p_i \in [0,1], N \in \mathbb{N} \tag{3.31}$$

is valid for $N = 2$ (and of course for $c = 2$).

(It can be noted that the validity of (3.31) for $N = 1$ is rather trivial)

Our objective shall be to prove the Shannon's entropy by mathematical induction with respect to N.

For performing the induction step, i.e. to show that (3.31) is valid for the natural number $N+1$ at the place of N, we start with the very fact that $H(p_1, p_2, \ldots, p_N, p_{N+1}) = H(p_1, p_{N+1}, p_2, p_3, \ldots, p_N)$ with subject to the usage of the postulate (I) with regard to the symmetry.

After this, we use the postulate (IV) to arrive at

$$H(p_1, p_{N+1}, p_2, p_3, \ldots, p_N) = H(p_1 + p_{N+1}, p_2, p_3, \ldots, p_N) + (p_1 + p_{N+1})H\left(\frac{p_1}{p_1+p_{N+1}}, \frac{p_{N+1}}{p_1+p_{N+1}}\right) \tag{3.73}$$

Now, we see that

- by assumption, (3.31) is valid for N probability elements implies

$$H(p_1+p_{N+1}, p_2, p_3, \ldots, p_N) = -(p_1+p_{N+1})\log_2(p_1+p_{N+1}) - \sum_{i=2}^{N} p_i \log_2 p_i \tag{3.74}$$

- by using (3.72), we get

$$H\left(\frac{p_1}{p_1+p_{N+1}}, \frac{p_{N+1}}{p_1+p_{N+1}}\right)$$

$$= -\frac{p_1}{p_1+p_{N+1}}\log_2\left(\frac{p_1}{p_1+p_{N+1}}\right) - \frac{p_{N+1}}{p_1+p_{N+1}}\log_2\left(\frac{p_{N+1}}{p_1+p_{N+1}}\right) \tag{3.75}$$

Therefore, by applying (3.74) and (3.75) on (3.73), we get

$$\begin{aligned} H(p_1, p_2, \ldots, p_N, p_{N+1}) &= H(p_1, p_{N+1}, p_2, p_3, \ldots, p_N) \\ &= -(p_1+p_{N+1})\log_2(p_1+p_{N+1}) - \sum_{i=2}^{N} p_i \log_2 p_i \\ &\quad - p_1 \log_2\left(\frac{p_1}{p_1+p_{N+1}}\right) - p_{N+1}\log_2\left(\frac{p_{N+1}}{p_1+p_{N+1}}\right) \\ &= -\sum_{i=2}^{N} p_i \log_2 p_i - p_1 \log_2 p_1 - p_{N+1}\log_2 p_{N+1} \\ &= \sum_{i=1}^{N+1} p_i \log_2\left(\frac{1}{p_i}\right) \end{aligned} \tag{3.76}$$

and this proves the Shannon's entropy for $N+1$ probability elements and hence the required ultimate proof of the Shannon's entropy (3.31) for $c = 2$.

In order to cope with the postulate (III), the logarithmic base of the logarithmic expression in (3.31) is generalized to any real value of c with $c > 1$. This completes the **generalized proof** of the expression (3.31), i.e. the proof of the **theorem 3.2.2**. □

3.2.4 Role of Shannon's four postulates

Since we have used the Shannon's postulates, namely postulates (I), (II), (III) and (IV), for deriving the Shannon's Entropy, it is however important for us to discuss, why these postulates have been at all used or precisely, which individual roles have each of these postulates have played. Before we go ahead, we need to state an important thing: The entropy of a probability distribution can also be interpreted as the **amount of information** contained in the probability distribution (or rather the **information content of the probability distribution**). We shall discuss this concept of information content in the subsequent subsection briefly. We shall discuss the importance of each of the Shannon's postulates one by one as follows:

1. For a discrete probability distribution with a finite support defined by the probabilities p_i, $i \in \{1, 2, \ldots, N\}$, we all know that the probability distribution is **not** changed, if the presentation of the **finite** sequential order of p_i is changed, i.e. if the order of presentation of the values p_i is changed. Therefore, since the measure of entropy is referred exclusively to the probability distribution and not the aforesaid order, we principally need to establish the very idea that the Shannon's entropy should to be independent of this aforesaid order. In other words, Shannon's entropy has to be a symmetric function of the arguments p_i, thereby the role of the postulate (I) is established.

2. The continuity of the entropy function $H(p, 1-p)$, $p \in [0, 1]$ (with the only exception of $H(p, 1-p)$ being removable discontinuous at $p = 0$) postulated by the postulate (II) basically makes sure that the entropy of a probability distribution is principally **not allowed to be unbounded above** under no circumstances. Since $H(p_1, p_2, \ldots, p_N)$ for $N \geq 2$ with subject to $\sum_{i=1}^{N} p_i = 1$ can be made recursively related to $H(p, 1-p)$ for a certain value of p, $p \in [0, 1]$ under the postulate (IV), we can be assured that $H(p_1, p_2, \ldots, p_N)$ **must be bounded above** for **every** set of values of p_i, $p_i \in [0, 1]$ and $i \in \{1, 2, \ldots, N\}$. This **boundedness** enables us to derive a probability distribution with **maximum** entropy. Therefore, the role of the postulate (II) is clear.

3. The postulate (III) solely sets the base of the logarithm involved in the entropy function $H(p_1, p_2, \ldots, p_N)$.

4. The role of the postulate (IV) is described as follows: Let us suppose that an event $\mathcal{E}$ of a random experiment with its probability $P_{\mathcal{E}}$ can be decomposed into two mutually exclusive events $\mathcal{E}_1$ and $\mathcal{E}_2$ with their respective probabilities $P_{\mathcal{E}_1}$ and $P_{\mathcal{E}_2}$, such that $P_{\mathcal{E}} = P_{\mathcal{E}_1} + P_{\mathcal{E}_2}$. Then, if we are informed concretely about the probabilities of these two subsets of the event $\mathcal{E}$ (i.e of these subsets $\mathcal{E}_1$ and $\mathcal{E}_2$), the amount of information obtained as a result is equal to the information content of the probability distribution defined by the two probabilities $\frac{P_{\mathcal{E}_1}}{P_{\mathcal{E}}}$ and $\frac{P_{\mathcal{E}_2}}{P_{\mathcal{E}}}$ multiplied with the weight $P_{\mathcal{E}}$, which is thereby given as $P_{\mathcal{E}} H\left(\frac{P_{\mathcal{E}_1}}{P_{\mathcal{E}}}, \frac{P_{\mathcal{E}_2}}{P_{\mathcal{E}}}\right)$. Basically, the postulate (IV) describes the **weighting** of different information contents as well as the **additivity** of information contents simultaneously.

3.2.5 A property of a Shannon's postulate

The following Shannon's postulate (could be termed as the postulate (V) is referred to the page 2 of [25]):

Statement 3.2.5 (Postulate V). *The entropy turns out to be zero, if one of the probability elements happens to represent a certain event, i.e. if* $p_i = 1$ *and* $p_j = 0$ *for every* $j \neq i$*, then the entropy reduces to zero, i.e.* $H(E_N) = 0$

As an **important property** of this postulate (V), it can be easily **deduced** by using the postulates (I) and (IV).

Notably, at the **end of the step 1** of the proof of the Shannon's theorem 3.2.2, we have seen that the **zero probability elements** do not contribute to the Shannon's entropy.

With this, keeping the postulate (I) in mind with regard to the **symmetry of the Shannon's entropy expression** and by using the postulate (IV) successively, we get $H(1, \underbrace{0, \ldots, 0}_{N \text{ zeros}}) = NH(0,1)\,(=0)$ (by (3.35)). This is nothing, but the **proof** of the postulate (V) and thereby the desired important property is established.

3.2.6 Shannon's entropy as the information content

In this subsection, we shall **briefly** introduce the application of the Shannon's entropy on the information theory. In the information theory, the Shannon's entropy is interpreted as the information content of the data meant for the transmission of information. In this case, the logarithmic base is appropriately chosen as $c = 2$. As a matter of fact, this logarithmic base $c = 2$ refers to the very fact that every information as a message is to be encoded as a sequence of zeros and ones (referred to the page 540 of [33]).

Let C_N define the set of all the characters, namely $z_1, z_2, \ldots, z_N$ contained in the data to be transmitted with individual probabilities of occurrences $p_1, p_2, \ldots, p_N$ respectively. C_N is thereby called the source of data.

In that case, we need to state and prove the following theorem referred to the page 79 of [7]:

Theorem 3.2.3 (The information content of a single character). *The information content of the character z_i, $i = 1, 2, \ldots, N$, i.e. the number of digits necessary to encode the character z_i can be derived to be*

$$I_i = \log_2 \left(\frac{1}{p_i} \right) \tag{3.77}$$

*Proof of the **theorem 3.2.3**.* The proof of the expression of the information content defined by (3.77) is rather simple and takes only few steps.

Let us start with the consideration of the independence of the events of occurrences of two characters z_i and z_j for $i \neq j$.

Accordingly, $p_{i,j} = p_i p_j$ and because the information contained jointly in both z_i and z_j is equal to the sum of the information contents of z_i and z_j individually, we have $I_{i,j} = I_i + I_j$, which can only bring us to $I_i = k \log_2(p_i)$, such that k is an arbitrary constant.

Now, under the consideration of a special case for $N = 2$, the characters (namely digits) 0 and 1 with their individual probabilities of occurrences $p_1 = 0.5$ and $p_2 = 0.5$ do correspond each of them to an information content of 1 bit only.

This leads to $1 = k \log_2(p_i)$, for $i = 1, 2$ and thereby giving $k = -1$.

Thus, $I_i = k \log_2(p_i) = \log_2\left(\frac{1}{p_i}\right)$, which proves the **theorem 3.2.3**. □

Definition 3.2.1 (Information content of a source code, the Shannon's formula). *By generalizing, the* ***information content***[3] *of the entire source of data denoted by E_N is basically the expected value of I_i. It is given by the weighted arithmetic mean of all the information contents I_i with p_is as weights, namely*

$$I_{E_N} = \sum_{i=1}^{N} p_i \log_2\left(\frac{1}{p_i}\right) \tag{3.78}$$

which gives us the ***average number of bits*** *necessary to encode each of the elements of E_N (referred to the page 80 of [7]).*

This information content of the aforesaid data source defined by (3.78) is also called (referred to the page 546 [33]) ***Shannon's formula****.*

Definition 3.2.2 (Hartley's formula). *As a special case for $p_i = \frac{1}{N}$ for every $i = 1, 2, \ldots, N$, the Shannon's formula reduces to $I_{E_N} = \log_2 N$, which is known as* ***Hartley's formula*** *(referred to the page 547 of [33]) and gives the number of bits necessary to code every character z_i, $i = 1, 2, \ldots, N$.*

In the information theory, this figure $\log_2 N$ is also interpreted as the ***number of binary decisions***[4] *(referred to the page 78 of [7]).*

Understandably, $\log_2 N$ is the ***maximum value*** *of I_{E_N}, simply because the entropy expressed by $\log_2 N$ of the discrete uniform probability distribution is maximum.*

A well known coding technique of the aforesaid characters is called the **procedure of Shannon and Fano**, which we shall state briefly as follows:

Proposition 3.2.1 (Coding procedure of Shannon and Fano). *The procedure (referred to the page 83 of [7]) says that the number of bits denoted by S_i needed to code the character z_i is determined by the constraint*

$$\log_2\left(\frac{1}{p_i}\right) \leq S_i \leq \log_2\left(\frac{1}{p_i}\right) + 1 \tag{3.79}$$

Remark 3.2.1 (The concluding remark). *We can well see that the Shannon's entropy has a* ***good usage*** *in the information theory.*

---

[3] The German word for information content is **Informationsgehalt**, referred to the page 80 of [7].

[4] The German word for this figure $\log_2 N$ is **Entscheidungsgehalt**.

3.3 The principle of maximum entropy

Probabilities are used to cope with the real aspect of randomness. However, randomness is only one source of uncertainty. The other source of uncertainty, which is even more severe, is the ignorance of the initial condition $\{d_Y\}$, which not only refers to the ignorant (parameter) space, but also to the structure of randomness. In the preceding section, the random structure as a function of $\{d_Y\}$ was handled by the principle of minimum information, by means of which the necessary amount of information in a given situation can be specified.

It has to be noted, that the minimum information principle determines the amount of qualitative information, which must be known in a given situation. These qualitative facts are gathered from certain experimental results (say from the drawing of samples).

In the year 1957, Edwin Jaynes investigated a somewhat related problem of selection of an appropriate probability distribution with subject to a given information about the initial conditions. It has to be noted, that **there is a difference between Jaynes's problem and our problem**. The minimum information principle answers the question of the **necessary amount of information**, whereas Jaynes starts with **a given amount of information**. Jaynes solved the problem based on the concept of stochastic entropy. For a given amount of information, he developed a selection principle for the probability distribution called *the principle of maximum entropy*.

According to the maximum entropy principle (referred to the page 1 of [25]), given some partial knowledge about Y (i.e. for eg. knowledge about the moments of Y), we should choose that particular probability distribution of Y, which is consistent or compatible with the given knowledge, but has otherwise maximum uncertainty associated with it.

Thus, the principle of maximum entropy starts with the assumption that only the confirmed knowledge can be used to develop a model, i.e., the probability distribution $\mathbf{P}_{Y|\{d_Y\}}$. Any **unconfirmed knowledge** used to build a stochastic model would lead to an **unspecified risk**. Any risk must have been generated by uncertainty and, therefore, in a given situation with only partial knowledge about the initial condition, that particular probability distribution should be selected which is compatible with the available knowledge

and exhibits the maximum stochastic uncertainty. Any different selection of $\mathbf{P}_{Y|\{d_Y\}}$ would be tantamount to the assumption of some unconfirmed knowledge resulting in an unknown risk (i.e. the risk that is not easily quantifiable). The maximum entropy probability measures (MEP-probability measures) as well as the maximum entropy probability densities are obtained by solving a constraint optimization problem.

Before we proceed, we would like to revisit the important (already stated) concepts of **type** and **information**, which are referred to a probability distribution of Y:

- The (specified) type of a probability distribution is described (given) by the number of extremal points of the probability distribution. All the probability distributions of a specified type are therefore qualitatively of the same type (or sort).
- The information needed to construct (or select) a probability distribution is described (given) by
  - the support of the probability distribution (absolutely necessary)
  - a certain number of moments, say m, $m \geq 0$ (no moments are necessary for a constant probability distribution)

3.3.1 The maximum entropy probability distribution

The derived statement giving the maximum entropy probability distribution with the help of **Kullback-Leibler divergence** (or **relative entropy**) happens to be the **theorem 12.1.1** given in the page 410 of the book [11] that has been published in the year **2006**. This particular derivation is given in the **appendix A.1** for the reader's ready reference.

In this subsection, we shall give our own derivation in a skillful manner. However, this derivation **cannot** be regarded as a complicated one and can be given by using **variously different** mathematical means. It has to be notably stated that I derived this very expression **completely in my own way** by using certain **natural logarithmic** properties in the year **2002**. The derivation is hereby given step by step as follows:

Let Ω be a **bounded Borel subset** of $\mathbb{R}$, $\mathcal{A}$ be a suitably chosen σ- algebra on Ω and ν be a σ- **finite measure** on $(\Omega, \mathcal{A})$, viz. $\nu(\Omega) < \infty$.

Let $\mathcal{F}(m)$ be the set of all the probability distributions $\mathbf{P}_{Y|\{d_Y\}}$ on $(\Omega, \mathcal{A})$, such that each of the elements of $\mathcal{F}(m)$, namely $\mathbf{P}_{Y|\{d_Y\}}$ possesses a density $\mathbf{f}_{Y|\{d_Y\}}$ with respect to ν and has the k th moment equal to a fixed number $\mu_Y^{(k)}$, for $k = 0, 1, 2, \ldots, m$, i.e.

$$\int_\Omega y^k \mathbf{P}_{Y|\{d_Y\}}(dy) = \int_\Omega y^k \mathbf{f}_{Y|\{d_Y\}}(y)\, \nu(dy) = \mu_Y^{(k)}, \;\; k = 0, 1, 2, \ldots, m$$

Basically, $\mu_Y^{(1)}, \mu_Y^{(2)} \ldots, \mu_Y^{(m)}$ are the numerical values of the first m moments of Y representing the available knowledge denoted by $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)} \ldots, \mu_Y^{(m)})$. Notably, $\mu_Y^{(0)} = 1$.

The entropy of $\mathbf{P}_{Y|\{d_Y\}}$ is hereby denoted by

$$H(\mathbf{P}_{Y|\{d_Y\}}) = -\int_\Omega \mathbf{f}_{Y|\{d_Y\}}(y) \log\left(\mathbf{f}_{Y|\{d_Y\}}(y)\right)\, \nu(dy) \tag{3.80}$$

which is invariant with respect to the dominating measure ν.

The maximum entropy principle says that $\mathbf{P}_{Y|\{d_Y\}}$ should be selected having maximum entropy and meeting the moment requirements simultaneously. Thus, we arrive at the following theorem:

Theorem 3.3.1 (Theorem of maximum entropy). *Corresponding to arbitrarily chosen real valued constants $\lambda_1, \lambda_2, \ldots, \lambda_m$, we set*

$$f_{Y|\{d_Y\}}(y) = e^{\sum_{k=0}^{m} \lambda_k y^k}, \;\; y \in \Omega \tag{3.81}$$

such that

$$e^{\lambda_0} = \frac{1}{\int_\Omega e^{\sum_{k=1}^{m} \lambda_k y^k}\, \nu(dy)}$$

where Ω is a bounded Borel subset of $\mathbb{R}$ and ν is a σ- finite measure on the Borel σ-algebra of Ω. Moreover, we define

$$\int_\Omega y^k f_{Y|\{d_Y\}}(y)\, \nu(dy) = \mu_Y^{(k)}, \;\; k = 0, 1, 2, \ldots, m;\; \mu_Y^{(0)} = 1 \tag{3.82}$$

Then the probability distribution $\mathbf{P}_{Y|\{d_Y\}}^{MEP}$ *possessing the* ν*- density* $f_{Y|\{d_Y\}}$ *has the maximum entropy among all the elements of the set* $\mathcal{F}(m)$*. This class* $\mathcal{F}(m)$ *of probability distributions, with subject to the fulfillment of*

$$\int_{\Omega} y^k \mathbf{f}_{Y|\{d_Y\}}(y)\, \nu(dy) = \mu_Y^{(k)}, k = 0, 1, 2, \ldots, m;\ \mu_Y^{(0)} = 1 \tag{3.83}$$

can be defined as

$$\mathcal{F}(m) = \left\{ \mathbf{f}_{Y|\{d_Y\}} : \Omega \to [0, \infty) \middle| \mathbf{f}_{Y|\{d_Y\}} \text{ is Borel measurable under (3.83)} \right\}$$

Consequently, $\mathbf{P}_{Y|\{d_Y\}}^{MEP} \in \mathcal{F}(m)$ *and* $H\left(\mathbf{P}_{Y|\{d_Y\}}^{MEP}\right) = \sup_{\mathbf{P}_{Y|\{d_Y\}} \in \mathcal{F}(m)} H\left(\mathbf{P}_{Y|\{d_Y\}}\right)$ *hold good. Moreover,* $H\left(\mathbf{P}_{Y|\{d_Y\}}^{MEP}\right) = -\sum_{k=0}^{m} \lambda_k \mu_Y^{(k)}$.

Proof of the ***theorem 3.3.1****.* Clearly, $\mathbf{P}_{Y|\{d_Y\}}^{MEP} \in \mathcal{F}(m)$.

For any arbitrarily chosen $\mathbf{P}_{Y|\{d_Y\}} \in \mathcal{F}(m)$ with a ν- density $\mathbf{f}_{Y|\{d_Y\}}$, we shall have to solve the following constraint optimization problem:

With subject to the fulfillment of the constraints (3.83), the expression (3.84) can be **globally** maximized. This maximization is equivalent to the maximization of the entropy $H(\mathbf{P}_{Y|\{d_Y\}})$. Here,

$$H(\mathbf{P}_{Y|\{d_Y\}}) + \lambda_0 + 1 + \lambda_1 \mu_Y^{(1)} + \lambda_2 \mu_Y^{(2)} + \ldots + \lambda_m \mu_Y^{(m)} \tag{3.84}$$

where the coefficients λ_i and the moments $\mu_Y^{(i)}$ for $i \in \{0, 1, 2, \ldots, m\}$ are described by (3.81) and (3.83) respectively.

Starting with (3.84) we get

$$
\begin{aligned}
& H\left(\mathbf{P}_{Y|\{d_Y\}}\right) + \lambda_0 + 1 + \lambda_1 \mu_Y^{(1)} + \ldots + \lambda_m \mu_Y^{(m)} \\
= & \int\limits_{\Omega} \mathbf{f}_{Y|\{d_Y\}}(y) \log\left(\frac{1}{\mathbf{f}_{Y|\{d_Y\}}(y)}\right) \nu(dy) \\
& + (\lambda_0 + 1) \int\limits_{\Omega} \mathbf{f}_{Y|\{d_Y\}}(y)\, \nu(dy) \\
& + \lambda_1 \int\limits_{\Omega} y \mathbf{f}_{Y|\{d_Y\}}(y)\, \nu(dy) + \ldots + \lambda_m \int\limits_{\Omega} y^m \mathbf{f}_{Y|\{d_Y\}}(y)\, \nu(dy) \\
= & \int\limits_{\Omega} \mathbf{f}_{Y|\{d_Y\}}(y) \left[\log\left(\frac{1}{\mathbf{f}_{Y|\{d_Y\}}(y)}\right) + \log\left(e^{\lambda_0+1}\right) + \log\left(e^{\lambda_1 y}\right)\right. \\
& \left. + \ldots + \log\left(e^{\lambda_m y^m}\right)\right] \nu(dy) \\
= & \int\limits_{\Omega} \mathbf{f}_{Y|\{d_Y\}}(y) \log\left(\frac{e^{\lambda_0+1+\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}}{\mathbf{f}_{Y|\{d_Y\}}(y)}\right) \nu(dy) \\
= & \int\limits_{\Omega} \mathbf{f}_{Y|\{d_Y\}}(y) \log\left(\frac{e^{1+\sum\limits_{i=0}^{m} \lambda_i y^i}}{\mathbf{f}_{Y|\{d_Y\}}(y)}\right) \nu(dy) \\
= & \int\limits_{\Omega} e^{1+\sum\limits_{i=0}^{m} \lambda_i y^i} \left[\frac{\mathbf{f}_{Y|\{d_Y\}}(y)}{e^{1+\sum\limits_{i=0}^{m} \lambda_i y^i}} \log\left(\frac{e^{1+\sum\limits_{i=0}^{m} \lambda_i y^i}}{\mathbf{f}_{Y|\{d_Y\}}(y)}\right)\right] \nu(dy)
\end{aligned}
$$

By the plain and simple fact that the function $x \log\left(\frac{1}{x}\right)$ of x having the domain of definition described by $0 < x \leq 1$ has its **global maximum value** $e^{-1} \log\left(\frac{1}{e^{-1}}\right) = e^{-1}$ at the point $x = e^{-1}$, the **least upper bound** (l.u.b.) for the above expression within the **first brackets** is obtained, thereby yielding the following result:

$$
\begin{aligned}
& \int\limits_{\Omega} e^{1+\sum\limits_{i=0}^{m} \lambda_i y^i} \left[\frac{\mathbf{f}_{Y|\{d_Y\}}(y)}{e^{1+\sum\limits_{i=0}^{m} \lambda_i y^i}} \log\left(\frac{e^{1+\sum\limits_{i=0}^{m} \lambda_i y^i}}{\mathbf{f}_{Y|\{d_Y\}}(y)}\right)\right] \nu(dy) \\
& \leq \int\limits_{\Omega} e^{1+\sum\limits_{i=0}^{m} \lambda_i y^i} \left[e^{-1}\right] \nu(dy) = \int\limits_{\Omega} e^{\sum\limits_{i=0}^{m} \lambda_i y_j^i} \nu(dy)
\end{aligned}
\tag{3.85}
$$

The l.u.b. on the right hand side of (3.85) is actually adopted at the point, where

$$\frac{\mathbf{f}_{Y|\{d_Y\}}(y)}{e^{1+\sum\limits_{i=0}^{m}\lambda_i y^i}} = e^{-1} \tag{3.86}$$

holds. From (3.86) the functional form of the probability density of the MEP-probability distribution is obtained at the point, when

$$\mathbf{f}_{Y|\{d_Y\}}(y) = f_{Y|\{d_Y\}}(y) = e^{\sum\limits_{i=0}^{m}\lambda_i y^i} \tag{3.87}$$

With (3.87), it is therefore been shown that on applying the maximum entropy principle to the case where the numerical values of the first m moments of $Y|\{d_Y\}$ are exactly known, the desired minimum information probability distribution is obtained. Each minimum information (i.e. the minimum available information necessary to construct the probability distribution of the desired type) probability distribution is an element of a given assigned co-domain of $\mathcal{P}$. Each such co-domain defines a distribution family that has been discussed previously.

Because of the very fact that $f_{Y|\{d_Y\}}$ given by (3.87) gives a probability distribution, by the derived result (3.85) as a result of the **global maximization** of the expression (3.84), we get

$$H\left(\mathbf{P}^{MEP}_{Y|\{d_Y\}}\right)+\lambda_0+1+\lambda_1\mu_Y^{(1)}+\lambda_2\mu_Y^{(2)}+\ldots+\lambda_m\mu_Y^{(m)} = \int\limits_{\Omega} e^{\sum\limits_{i=0}^{m}\lambda_i y^i}\,dy = 1 \tag{3.88}$$

implying that the maximum stochastic entropy is given by:

$$H\left(\mathbf{P}^{MEP}_{Y|\{d_Y\}}\right) = \Gamma(\lambda_1,\lambda_2,\ldots,\lambda_m) = -\sum_{i=0}^{m}\lambda_i\mu_Y^{(i)} \tag{3.89}$$

Thus, the entropy of the $(m-1)$- extremal minimum information probability distribution is therefore given by the functional given by (3.89). Clearly, $H\left(\mathbf{P}_{Y|\{d_Y\}}\right) \leq H\left(\mathbf{P}^{MEP}_{Y|\{d_Y\}}\right)$ holds and this **proves our theorem 3.3.1**. □

Remark 3.3.1. *In contrast to the* ***maximum entropy principle****, the* ***minimum information principle*** *guarantees that the selected probability distribution, which describes, in general, the given situation,* ***sufficiently well****.*

As **special cases** of the **theorem 3.3.1**, the following two corollaries are discussed to consider the **discrete** and **continuous** cases of Y **individually**:

Corollary 3.3.1 (Y is discrete). *By taking $\Omega = \{y_1, y_2, \ldots, y_N\}$ and ν to be a **counting measure** on Ω, i.e. $\nu(\{y_j\}) = j \leq N$, the corollary is described as follows:*

*Discussion of the **corollary 3.3.1**.* Let $\mathcal{F}_D(m)$ be the set of all the probability mass functions with the support $\{y_1, y_2, \ldots, y_N\}$, such that $y_1 < y_2 < \ldots < y_N$ and the k th moment of each of the elements of $\mathcal{F}_D(m)$ is a fixed number $\mu_Y^{(k)}$, for $k = 0, 1, 2, \ldots, m$. Symbolically,

$$\mathcal{F}_D(m) = \left\{ (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N) \in [0,1]^N : \sum_{j=1}^{N} y_j^k \mathbf{p}_j = \mu_Y^{(k)}, 0 \leq k \leq m, \right\}$$

with $\mu_Y^{(0)} = 1$

The probability mass function chosen from $\mathcal{F}_D(m)$, which maximizes the entropy $H(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N) = -\sum_{j=1}^{N} \mathbf{p}_j \log(\mathbf{p}_j)$ at the point, when

$$\mathbf{p}_j = p_j = P_{Y|\{d_Y\}}(\{y_j\}) = e^{\sum_{k=0}^{m} \lambda_k y_j^k}, \;\; 1 \leq j \leq N \tag{3.90}$$

which describes the **discrete case** of Y.

The entropy of the maximum entropy discrete probability distribution of Y is accordingly given as

$$H\left(P_{Y|\{d_Y\}}\right) = -\sum_{k=0}^{m} \lambda_k \mu_Y^{(k)} \tag{3.91}$$

and this completes the **corollary 3.3.1**. □

Corollary 3.3.2 (Y is continuous). *By taking $\Omega = [a, b]$ to be a closed interval in $\mathbb{R}$ and ν to be a **Lebesgue measure** on $[a, b]$, the corollary is described as follows:*

*Discussion of the **corollary 3.3.2**.* Let $\mathcal{F}_C(m)$ be the set of all the probability density functions with the support $[a, b]$, such that the k th moment of

each of the elements of $\mathcal{F}_C(m)$ is a fixed number $\mu_Y^{(k)}$, for $k = 0, 1, 2, \ldots, m$. Symbolically,

$$\mathcal{F}_C(m) = \left\{ \mathbf{f}_{Y|\{d_Y\}} : [a,b] \to [0,\infty) : \int_a^b y^k \mathbf{f}_{Y|\{d_Y\}}(y) dy = \mu_Y^{(k)}, 0 \leq k \leq m \right\}$$

with $\mu_Y^{(0)} = 1$.

Now, by rewriting the statement of the theorem in this special case, we state:

With subject to the existence of the real valued constants $\lambda_0, \lambda_1, \lambda_2 \ldots, \lambda_m$, if

$$f_{Y|\{d_Y\}}(y) = e^{\sum_{k=0}^{m} \lambda_k y^k}, \;\; a \leq y \leq b \tag{3.92}$$

fulfills the condition

$$\int_a^b y^k f_{Y|\{d_Y\}}(y)\, dy = \mu_Y^{(k)}, \;\; k = 0, 1, 2, \ldots, m;\; \mu_Y^{(0)} = 1$$

then $f_{Y|\{d_Y\}}$ maximizes the entropy within $\mathcal{F}_C(m)$ and this describes the **continuous case** of Y (the special case for $a = 0$, $m = 2$ in this regard can be referred to the page 71 of [25]).

The entropy of the maximum entropy continuous probability distribution of Y is accordingly given as

$$H\left(f_{Y|\{d_Y\}}\right) = -\sum_{k=0}^{m} \lambda_k \mu_Y^{(k)} \tag{3.93}$$

by replacing the argument $P_{Y|\{d_Y\}}$ of H by the argument $f_{Y|\{d_Y\}}$ in this continuous case. This completes the **corollary 3.3.2** □

3.3.2 The Kullback-Leibler measure of deviation

With respect to a **fixedly given support**, namely $\{y_1, y_2, \ldots, y_N\}$ (where for a finite N number of support elements, $a = y_1 < y_2 < \ldots < y_N = b$)

in case Y is discrete or $[a, b]$ in case Y is continuous, the uniform probability distribution followed by Y has understandably the **maximum entropy** compared to any other probability distribution followed by Y.

The **deviation** (or **divergence**) of a probability distribution followed by Y **from** the uniform probability distribution followed by Y is measured by the **Kullback-Leibler divergence** (or in other words, relative entropy). This particular kind of deviation measured by the relative entropy is an **alternative way** of interpreting the entropy of the probability distribution of Y.

In our discussion, we shall select that particular probability distribution of Y, which is determinable by its first m moments, namely determinable by $d_Y = \left(\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}\right)$.

Without any loss of generality, let us take $a = 0$ and $b = 1$ in case of a **continuous** Y and $0 = y_1 < y_2 < \ldots < y_N = 1$ in case of a **discrete** Y. We shall discuss the **continuous case at first** and then the **discrete case** in form of the following two propositions.

Proposition 3.3.1 (The Kullback-Leibler deviation in the continuous case of Y). *The Kullback-Leibler deviation between the **maximum entropy** probability distribution of Y determined by $d_Y = \left(\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}\right)$ and the **uniform** probability distribution of Y, both having the exactly the same support $[0, 1]$, is **minimum** and is equal to $-H\left(f_{Y|\{d_Y\}}\right)$.*

*Proof of the **proposition 3.3.1**.* Here, if $f_{Y|\{d_U\}}$ is the probability density function of the uniform continuous probability distribution (d_U gives the moments of the uniform continuous distribution) and $\mathbf{f}_{Y|\{d_Y\}}$ be any other probability distribution belonging to $\mathcal{F}_C(m)$, then by Gibbs' inequality,

$$\begin{aligned} &\int_0^1 \mathbf{f}_{Y|\{d_Y\}}(y) \log\left(\frac{\mathbf{f}_{Y|\{d_Y\}}(y)}{1}\right) dy \geq 0 \\ \Leftrightarrow &H\left(\mathbf{f}_{Y|\{d_Y\}}\right) = \int_0^1 \mathbf{f}_{Y|\{d_Y\}}(y) \log\left(\frac{1}{\mathbf{f}_{Y|\{d_Y\}}(y)}\right) dy \leq 0 \end{aligned} \tag{3.94}$$

which clearly shows that the entropy of any continuous probability distribution with support $[0, 1]$, with the exception of the uniform continuous proba-

bility distribution is **always negative** (the entropy of the uniform continuous probability distribution being **zero**).

Therefore, with regard to the support $[0, 1]$ and any fixedly chosen d_Y, the entropy of the probability distribution described by the density $f_{Y|\{d_Y\}}$ of Y is **maximum** precisely means that the **Kullback-Leibler deviation** between $f_{Y|\{d_Y\}}$ and $f_{Y|\{d_U\}}$ is **understandably minimum** in comparison with **all other** probability distributions $\mathbf{f}_{Y|\{d_Y\}}$ belonging to $\mathcal{F}_C(m)$.

Thus, this Kullback-Leibler deviation between $\mathbf{f}_{Y|\{d_Y\}} = f_{Y|\{d_Y\}}$ and $f_{Y|\{d_U\}}$, namely

$$-H\left(f_{Y|\{d_Y\}}\right) = \int_0^1 f_{Y|\{d_Y\}}(y) \log\left(f_{Y|\{d_Y\}}(y)\right) dy$$

gives the **measure of the minimum deviation** of the probability distribution of Y determined by d_Y (i.e. determined by the first m moments), where the **probability density function** $f_{Y|\{d_Y\}}$ has been described by (3.92).

The probability distribution $f_{Y|\{d_U\}}$ having **zero** as the **entropy with respect to the support** $[0, 1]$ may be termed as the probability distribution representing the **complete unknownness**.

This completes the discussion as well as the proof of the **proposition 3.3.1**. □

The discussions about the discrete cases of Y is more or less the same, except the number of elements of the support denoted by N is of importance.

Proposition 3.3.2 (The Kullback-Leibler deviation in the discrete case of Y). *The Kullback-Leibler deviation between the* ***maximum entropy*** *probability distribution of Y determined by $d_Y = \left(\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}\right)$ and the* ***uniform*** *probability distribution of Y, both having the exactly the same support $\{y_1, y_2, \ldots, y_N\}$, is* ***minimum*** *and is equal to $-H\left(P_{Y|\{d_Y\}}\right) + \log N$.*

Proof of the ***proposition 3.3.2****.* If $P_{Y|\{d_U\}}$ is the probability mass function of the uniform discrete probability distribution (d_U gives the moments of the uniform discrete distribution) and $\mathbf{P}_{Y|\{d_Y\}}$ be any other probability distribution belonging to $\mathcal{F}_D(m)$ for any fixedly chosen N, then exactly by the **same**

argument as in the continuous case of Y, the Kullback - Leibler deviation between $\mathbf{P}_{Y|\{d_Y\}} = P_{Y|\{d_Y\}}$ and $P_{Y|\{d_U\}}$ is given by

$$\sum_{j=1}^{N} P_{Y|\{d_Y\}}(\{y_j\}) \log\left(\frac{P_{Y|\{d_Y\}}(\{y_j\})}{\frac{1}{N}}\right)$$
$$= \sum_{j=1}^{N} P_{Y|\{d_Y\}}(\{y_j\}) \log\left(P_{Y|\{d_Y\}}(\{y_j\})\right) + \log N = -H\left(P_{Y|\{d_Y\}}\right) + \log N$$

which gives the **measure of the minimum deviation** of the probability distribution of Y determined by d_Y (i.e. determined by the first m moments), where the **probability mass function** $P_{Y|\{d_Y\}}$ has been described by (3.90).

This completes the discussion as well as the proof of the **proposition 3.3.2**. □

Now, let us formulate the concept of complete information in the following way:

Remark 3.3.2 (The complete information). ***Conclusively****, in* ***this very particular sense****, the probability distribution of Y ($P_{Y|\{d_Y\}}$ in the discrete case of Y or $f_{Y|\{d_Y\}}$ in the continuous case of Y), which gives the* ***minimum deviation*** *(i.e. the* ***minimum Kullback - Leibler deviation****), contains the* ***complete information*** *of the existing but unknown probability distribution of Y determined by d_Y (i.e. the probability distribution determined by its first m moments).*

Before we draw this very subsection to a close, we would like to give a brief note about the **moments** and **supports** of the probability distributions of both Y and X for our future references:

Remark 3.3.3. *Regardless of whether the random variable Y is discrete or continuous, we shall express the i^{th} moment of Y, namely $\mu_Y^{(i)}$ for our references in the coming chapters and sections as*

$$\mu_Y^{(i)} = \frac{\int\limits_{\Omega} y^i e^{\sum\limits_{j=1}^{m} \lambda_j y^j} \nu(dy)}{\int\limits_{\Omega} e^{\sum\limits_{j=1}^{m} \lambda_j y^j} \nu(dy)} = \int\limits_{\Omega} y^i e^{\sum\limits_{j=0}^{m} \lambda_j y^j} \nu(dy)$$

by using the statements (3.81) and (3.82).

In particular, if Y is linearly transformed as $Y = y_1 + (y_N - y_1)X$ in the ***discrete case*** *or $Y = a + (b-a)X$ in the* ***continuous case****, then the i^{th} of the transformed random variable X, namely μ_i shall be expressible as*

$$\mu_i = \frac{\int\limits_{\Omega_X} x^i e^{\sum\limits_{j=1}^{m} \beta_j x^j} \nu_X(dx)}{\int\limits_{\Omega_X} e^{\sum\limits_{j=1}^{m} \beta_j x^j} \nu_X(dx)} = \int\limits_{\Omega_X} x^i e^{\sum\limits_{j=0}^{m} \beta_j x^j} \nu_X(dx)$$

where, as a result of this linear transformation, the transformations of Ω and ν are symbolized by Ω_X and ν_X respectively.

However, in course of our discussions, we shall use $\mathcal{X}_Y$ instead of Ω and $\mathcal{X}_X$ instead of Ω_X.

3.3.3 A preliminary statement pertaining to the characteristic properties of λ_i, $i \in \{0, 1, 2, \ldots, m\}$ values

In the subsequent subsection 3.3.4, we shall discuss about the **characteristic properties** of λ_i, $i \in \{0, 1, 2, \ldots, m\}$ values pertaining to both **discrete** and **continuous** cases of Y. We must state the following prior to the aforesaid discussions:

Statement 3.3.1 (The basic assumption). *The value of* ***m*** *is always* ***arbitrarily*** *but* ***fixedly*** *chosen in advance. Only with respect to this fixed choice of* ***m****, N is allowed to vary within the range $m + 1 \leq N < \infty$.*

In fact, the characteristic properties of λ_i, $i \in \{0, 1, 2, \ldots, m\}$ values are ascribed to the discussions about the ***exponential polynomial*** *probability distributions of Y having* ***at most*** *$m - 1$ extremes.*

Throughout the **subsection 3.3.4**, we shall go by the **statement 3.3.1**.

3.3.4 Characteristic properties of λ_i, $i \in \{0, 1, 2, \ldots, m\}$ values

Definition 3.3.1 (The desired property of a probability distribution). *The number of* ***extreme points*** *of the probability distribution of Y is the basic* ***desired*** *characteristic property for the* ***selection*** *of the probability distribution of Y.*

This property is exclusively determined by the λ_i, $i \in \{0, 1, 2, \ldots, m\}$ values.

Statement 3.3.2 (The basic characteristic property). *If N is made to be* ***extremely large****, then the random variable Y may be taken for a* ***continuous****, instead of being* ***discrete****. Even in cases, when Y is taken for a continuous, the λ_i, $i \in \{1, 2, \ldots, m\}$ values must be* ***kept unchanged****.*

Meaning of the ***statement 3.3.2****.* The term continuous probability distribution of the random variable $Y|\{d_Y\}$ means that the probability distribution of $Y|\{d_Y\}$ is describable by its probability density function rather than its probability mass function, if N happens to be extremely large, i.e. mathematically speaking, if $N \to \infty$. In other words, only if N happens to be **sufficiently large**, the probability distribution of $Y|\{d_Y\}$ could be taken for a continuous one instead of a discrete one. As a matter of fact, it is extremely important to state that the values of $\lambda_1, \lambda_2, \ldots, \lambda_m$ contained in the probability mass function of $Y|\{d_Y\}$ must remain unchanged if $Y|\{d_Y\}$ happens to be taken for a continuous random variable, simply because these λ_i values determine the essential characteristics of the probability distribution, especially the **local extrema**. The continuous probability distribution (of $Y|\{d_Y\}$) therefore should not lead to any change in the essential characteristics of the probability distribution.

This aforesaid **largeness** of N is therefore the **specific condition** for choosing the probability distribution of Y to be **continuous**, **instead of discrete**.

This ends the discussion of the **statement 3.3.2**. □

In view of the **definition 3.3.1** as well as the **statement 3.3.2**, we shall **assume** in this very subsection, that the λ_i, $i \in \{1, 2, \ldots, m\}$ values for both the **discrete** and the **continuous** cases of Y are exactly the **same**. By keeping this in mind, let us examine, how the moments of Y in both the discrete and the continuous cases behave.

Proposition 3.3.3 (Behavior of moments of Y). *If N is **sufficiently large** for any **fixedly** chosen $\boldsymbol{m}$, then in the **limiting sense**, the i^{th} moment ($i \in \{1, 2, \ldots, m\}$) of Y in the **continuous** case (denoted by $\mu_Y^{(i,C)}$) can be taken to be **equal** to the i^{th} moment of Y in the **discrete** case (denoted by $\mu_Y^{(i,D)}$).*

*Proof of the **proposition 3.3.3**.* Let us **restate** the the maximum entropy probability distribution of the random variable $Y|\{d_Y\}$ in the **discrete** case (in case of the **Shannon's entropy**) as well as in the **continuous** case (in case of the **differential entropy** [5]). For a **discrete** $Y|\{d_Y\}$, its **maximum entropy probability distribution** is given as

$$f_{Y|\{d_Y\}}(y_j) = e^{\sum_{i=0}^{m} \lambda_i y_j^i} \text{ for } j = 1, \ldots, N$$

the **value of its entropy** being $-\sum_{i=0}^{m} \lambda_i \mu_Y^{(i,D)}$, where each moment $\mu_Y^{(i,D)}$ symbolized for showing the discrete case is given as

$$\mu_Y^{(i,D)} = \frac{\sum_{j=1}^{N} y_j^i e^{\sum_{k=1}^{m} \lambda_k y_j^k}}{\sum_{j=1}^{N} e^{\sum_{k=1}^{m} \lambda_k y_j^k}}$$

and for a **continuous** $Y|\{d_Y\}$, its **maximum entropy probability distribution** is given as

$$f_{Y|\{d_Y\}}(y) = e^{\sum_{i=0}^{m} \lambda_i y^i} \text{ for } a \leq y \leq b$$

the **value of its entropy** being $-\sum_{i=0}^{m} \lambda_i \mu_Y^{(i,C)}$, where each moment $\mu_Y^{(i,C)}$ symbolized for showing the continuous case is given as

$$\mu_Y^{(i,C)} = \frac{\int_a^b y^i e^{\sum_{k=1}^{m} \lambda_k y^k} dy}{\int_a^b e^{\sum_{k=1}^{m} \lambda_k y^k} dy}$$

[5] the **definition** of the **differential entropy** is also referred to the page 44 of [25]

With this [6], for $\delta_{\max} = \max\limits_{1\leq j\leq N-1}(y_{j+1} - y_j) \to 0 \Leftrightarrow N \to \infty$,

$$\begin{aligned}
\lim_{N\to\infty} \mu_Y^{(i,D)} &= \lim_{N\to\infty} \left(\frac{\sum\limits_{j=1}^{N} y_j^i e^{\sum\limits_{k=1}^{m} \lambda_k y_j^k}}{\sum\limits_{j=1}^{N} e^{\sum\limits_{k=1}^{m} \lambda_k y_j^k}} \right) \\
&= \lim_{N\to\infty} \left(\frac{\sum\limits_{j=1}^{N} \delta_{\max} y_j^i e^{\sum\limits_{k=1}^{m} \lambda_k y_j^k}}{\sum\limits_{j=1}^{N} \delta_{\max} e^{\sum\limits_{k=1}^{m} \lambda_k y_j^k}} \right) \\
&= \frac{\lim\limits_{\delta_{\max}\to 0} \sum\limits_{j=1}^{N} \delta_{\max} y_j^i e^{\sum\limits_{k=1}^{m} \lambda_k y_j^k}}{\lim\limits_{\delta_{\max}\to 0} \sum\limits_{j=1}^{N} \delta_{\max} e^{\sum\limits_{k=1}^{m} \lambda_k y_j^k}} \text{ by keeping } a = y_1 \text{ and } b = y_N \text{ fixed} \\
&= \frac{\int\limits_a^b y^i e^{\sum\limits_{k=1}^{m} \lambda_k y^k} dy}{\int\limits_a^b e^{\sum\limits_{k=1}^{m} \lambda_k y^k} dy} = \mu_Y^{(i,C)} \qquad (3.95)
\end{aligned}$$

Thus, for a **sufficiently large** value of N, $\mu_Y^{(i,D)} \approx \mu_Y^{(i,C)}$ and this completes the proof of the **proposition 3.3.3**. □

Consequently, we immediately arrive at the following corollary:

Corollary 3.3.3. $-\sum\limits_{i=1}^{m} \lambda_i \mu_Y^{(i,D)} \approx -\sum\limits_{i=1}^{m} \lambda_i \mu_Y^{(i,C)}$ *for a sufficiently large* N.

Basically, for exactly the same values of $\lambda_1, \lambda_2, \ldots \lambda_m$, the values of the moments, namely $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots \mu_Y^{(m)}$ can be kept almost unchanged by increasing N arbitrarily, however with subject to the fixed values of $a = y_1$ and $y_N = b$.

[6] Notably, the differential entropy $\int\limits_a^b \mathbf{f}_{Y|\{d_Y\}}(y) \log\left(\frac{1}{\mathbf{f}_{Y|\{d_Y\}}(y)}\right) dy$ assumes its **global** maximum value, when $\mathbf{f}_{Y|\{d_Y\}}(y) = e^{\sum\limits_{i=0}^{m} \lambda_i y^i}$ for $a \leq y \leq b$.

But, as a matter of fact the value of λ_0 **cannot** be kept unchanged and therefore needs **careful investigation**.

Keeping this in mind, we arrive at the following proposition:

Proposition 3.3.4. ***Purely*** *for the sake of* ***simplicity*** *and* ***convenience****, we shall assume that* δ_j *values defined by* $\delta_j = y_{j+1} - y_j$, $j = 1, 2, \ldots, N-1$ *are* ***not significantly distinct****. In other words,* $\delta_j \approx \delta = \frac{b-a}{N-1}$ *for every* $j = 1, 2, \ldots, N-1$ ***always*** *holds for a* ***reasonably large*** N*.*

With this, let the values of λ_0 *in discrete and continuous cases be denoted by* $\lambda_0^{(D)}$ *and* $\lambda_0^{(C)}$ *respectively. Then, for a* ***sufficiently large*** *value of* N*,* $\lambda_0^{(C)}$ *differs from* $\lambda_0^{(D)}$ *approximately by the expression* $\log\left(\frac{b-a}{N-1}\right)$*, where* a *and* b *have their usual meanings.*

Proof of the ***proposition 3.3.4****.* Notably, if N is **not reasonably large**, then one should **not** opt for choosing $Y|\{d_Y\}$ to be a **continuous** random variable.

Therefore, with regard to the defined δ_j values, namely $\delta_j = y_{j+1} - y_j$, $j = 1, 2, \ldots, N-1$, together with $\delta = \frac{b-a}{N-1}$, we set

$$1 \approx \sum_{j=1}^{N} \delta_j e^{\lambda_0^{(C)} + \sum_{i=1}^{m} \lambda_i y_j^i} \approx \frac{b-a}{N-1} \sum_{j=1}^{N} e^{\lambda_0^{(C)} + \sum_{i=1}^{m} \lambda_i y_j^i} \tag{3.96}$$

for a **large** N, which on further simplification gives

$$\sum_{j=1}^{N} \delta_j e^{\sum_{i=1}^{m} \lambda_i y_j^i} \approx \frac{b-a}{N-1} \sum_{j=1}^{N} e^{\sum_{i=1}^{m} \lambda_i y_j^i} \tag{3.97}$$

after having taken $\delta_N = 0$. **Notably, the derivation of (3.95) assumed that** m **is not made to change with** N **for eg.** $m = N - 3$**. This meant,** m **has been assumed to be independent of** N.

So, by keeping this in mind, let us find out a concrete relationship between $\lambda_0^{(D)}$ and $\lambda_0^{(C)}$.

With subject to the very following fact

$$\lambda_0^{(D)} = -\log\left(\sum_{j=1}^{N} e^{\sum_{i=1}^{m} \lambda_i y_j^i}\right) \tag{3.98}$$

we get

$$\begin{aligned}\lambda_0^{(C)} &= -\log\left(\int_a^b e^{\sum_{i=1}^{m}\lambda_i y^i}\,dy\right)\\ &= -\log\left(\lim_{N\to\infty}\sum_{j=1}^{N}\delta_j e^{\sum_{i=1}^{m}\lambda_i y_j^i}\right)\\ &\approx -\log\left(\sum_{j=1}^{N}\delta_j e^{\sum_{i=1}^{m}\lambda_i y_j^i}\right)\\ &\approx -\log\left(\frac{b-a}{N-1}\sum_{j=1}^{N}e^{\sum_{i=1}^{m}\lambda_i y_j^i}\right) = -\log\left(\frac{b-a}{N-1}\right)+\lambda_0^{(D)}\end{aligned} \tag{3.99}$$

which means

$$\lambda_0^{(C)} \approx -\log\left(\frac{b-a}{N-1}\right)+\lambda_0^{(D)} \tag{3.100}$$

This completes the proof of the **proposition 3.3.4**. □

Purely for the sake of a possible clarity, we restate the immediately above statement in form of the following corollary:

Corollary 3.3.4. *The derived (3.100) basically means nothing different from the very fact that* $-\log\left(\frac{b-a}{N-1}\right)+\lambda_0^{(D)} \to \lambda_0^{(C)}$, *if* $N \to \infty$.

Consequently, there are certain corollaries, which relevantly arise:

Corollary 3.3.5 (The selection criterion). *The word selection is referred to the choice, whether to choose the* ***discrete*** *probability distribution of* Y *or rather the* ***continuous*** *of the same.*

If the ***positive*** *real number*

$$\left|\lambda_0^{(C)}+\log\left(\frac{b-a}{N-1}\right)-\lambda_0^{(D)}\right| \tag{3.101}$$

is ***sufficiently small*** *(because of the* ***largeness*** *of* N*), then the* ***continuous*** *probability distribution of* Y *with* $\lambda_0^{(C)}$, λ_i, $i \in \{1,2,\ldots,m\}$ *could be selected for usages,* ***instead of*** *the* ***discrete*** *probability distribution of* Y *with* $\lambda_0^{(D)}$, λ_i, $i \in \{1,2,\ldots,m\}$, *provided the* λ_i, $i \in \{1,2,\ldots,m\}$ *values for both continuous and the discrete cases of* Y *are* ***practically*** *the same.*

Before we go ahead to establish the next corollary, we need to state a very important theorem that connects the Shannon's entropy and the differential entropy[7], which states that

Theorem 3.3.2 (Theorem of Shannon's and the differential entropy). *If the density $\mathbf{f}_Y(y)$ of the **continuous** random variable Y is Riemann integrable, such that $\mathbf{f}_Y(\mathbf{y}_j)\,\delta$, $j \in \mathbb{Z}$ is taken for a **probability element** in case Y is taken for a **discrete**, then*

***the Shannon's entropy** $+\log(\delta) \rightarrow$ **the differential entropy, as** $\delta \rightarrow 0$ ($\delta > 0$).*

This theorem is referred to the **theorem 8.3.1** in the page 248 of [11].
With this, we are in a position to come to the next corollary, which is given as follows:

Corollary 3.3.6 (Validity of the theorem 3.3.2 in our case). *We know that the Shannon's entropy is equal to $-\lambda_0^{(D)} - \sum_{i=1}^{m} \lambda_i \mu_Y^{(i,D)}$ and the differential entropy is equal to $-\lambda_0^{(C)} - \sum_{i=1}^{m} \lambda_i \mu_Y^{(i,C)}$. For a **large** N, $\mu_Y^{(i,D)} \approx \mu_Y^{(i,C)}$ holds.*

*So, by combining the preceding **corollaries 3.3.3** and **3.3.4**, we get*

$$-(-\log\left(\tfrac{b-a}{N-1}\right) + \lambda_0^{(D)}) - \sum_{i=1}^{m} \lambda_i \mu_Y^{(i,D)} \rightarrow -(\lambda_0^{(C)}) - \sum_{i=1}^{m} \lambda_i \mu_Y^{(i,C)} \text{ as } N \rightarrow \infty$$

i.e.

$$\underbrace{\log\left(\frac{b-a}{N-1}\right)}_{=\log(\delta)} - \lambda_0^{(D)} - \sum_{i=1}^{m} \lambda_i \mu_Y^{(i,D)} \rightarrow -\lambda_0^{(C)} - \sum_{i=1}^{m} \lambda_i \mu_Y^{(i,C)} \text{ as } N \rightarrow \infty$$

*since in our case, $\delta = \frac{b-a}{N-1}$. **Notably**, if the elements $y_1, y_2, \ldots, y_N$ of the support $\mathcal{X}_Y$ are **equidistant**, then $\delta_j = \delta = \frac{b-a}{N-1}$, $j \in \{1, 2, \ldots, N-1\}$.*

*Hence, the **theorem 3.3.2** gets verified here.*

Importantly, in the next corollary we shall see trivially, how exactly $\lambda_0^{(D)}$ behaves with $N \rightarrow \infty$.

Corollary 3.3.7 ($\lambda_0^{(D)} \rightarrow -\infty$ as $N \rightarrow \infty$). *By rewriting (3.100), that is*

$$\lambda_0^{(C)} + \log\left(\frac{b-a}{N-1}\right) \approx \lambda_0^{(D)} \tag{3.102}$$

[7]Notably, the usage of the term **differential entropy** has **also** been referred to the page 247 of [11]

we can easily see that, since $\lambda_0^{(C)}$ *is* ***independent*** *of* N*, then if* $N \to \infty$*, then* $\lambda_0^{(D)}$ *understandably tends to* $-\infty$*.*

That is, $N \to \infty \implies \lambda_0^{(D)} \to -\infty$ *and this establishes this corollary.*

This brings us to the following important statement regarding higher values of N:

Statement 3.3.3 (High values of N). *For a very high value of* N*, say 10000, if we still decide to stick to the discrete case,* $\lambda_0^{(D)}$ *is ought to be a negative value of a very high magnitude, which evidently means the following:*

1. *Entropy of the discrete probability distribution under consideration, namely* $-\lambda_0^{(D)} - \sum_{i=1}^{m} \lambda_i \mu_Y^{(i)}$*, is expectedly very high, because of the positive value of* $-\lambda_0^{(D)}$ *of high magnitude.*

2. *The individual probabilities, namely* $e^{\lambda_0^{(D)} + \sum_{i=1}^{m} \lambda_i y_j^i}$ *become undoubtedly individually smaller.*

3. *Even for large values of* N*, the value* $\lambda_0^{(C)} = -\log(\frac{b-a}{N-1}) + \lambda_0^{(D)}$*, is not really expected to be a value of a very high magnitude as* $\log(N-1)$ *(*> 0*) neutralizes the magnitude of* $\lambda_0^{(D)}$ *to a considerable extent.*

Conclusively, in this very subsection, we have achieved the following:

- If $Y|\{d_Y\}$ is to be continuous, there does not seem to be any harm in stating its **differential** entropy defined by the pure and simple Riemann integral $\int_a^b f_{Y|\{d_Y\}}(y) \log\left(\frac{1}{f_{Y|\{d_Y\}}(y)}\right) dy = -\lambda_0^{(C)} - \sum_{i=1}^{m} \lambda_i \mu_Y^{(i)}$ to be the **information content** of its probability distribution, its probability density function being given by $f_{Y|\{d_Y\}}(y) = e^{\lambda_0^{(C)} + \sum_{i=1}^{m} \lambda_i y^i}$ for $a \leq y \leq b$.

- We have established a clear **relationship** between the λ_i values for $i \in \{0, 1, 2, \ldots, m\}$ of the probability density function of $Y|\{d_Y\}$ (in case of a continuous $Y|\{d_Y\}$) and the same of the probability mass function of $Y|\{d_Y\}$ (in case of a discrete $Y|\{d_Y\}$) **with regard to the largeness** of N.

- With subject to the largeness of N, the user may decide, whether to take the random variable $Y|\{d_Y\}$ for **discrete** or rather **continuous**.

3.3.5 Supporting numerical examples

If N is sufficiently high, we shall illustrate by two numerical examples that, corresponding to the fixedly chosen $\mu_Y^{(i,D)} = \mu_Y^{(i,C)}$, $i \in \{1, 2, \ldots, m\}$, we have $\lambda_i^{(D)} \to \lambda_i^{(C)}$ as $N \to \infty$ for every $i \in \{1, 2, \ldots, m\}$, but for a fixed m.

These numerical examples have been set with the help of certain developed software programs belonging to this thesis, such that the support of a discrete probability distribution is always assumed to be the elements in **arithmetic progression** for **simplicity's sake**. The first example is an example of **unimodal** probability distribution of Y and the second example is of **bathtub** probability distribution of Y. These are illustrated as follows:

Example 3.3.1. *Here, we take* $a = -20$, $b = 20$, $m = 2$, $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 16.5741$.

In accordance with the above data, the continuous approximated density function is given by

$$f_{Y|\{d_Y\}} = e^{-2.734764547565833+0.5721739130434783y-0.08695652173913045y^2}, \quad -20 \leq y \leq 20.$$

So, corresponding to the discrete support

1. $\{y_1 = -20, -15, -10, -5, 0, 5, 10, 15, 20 = y_9\}$ *(here,* $N = 9$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 16.5741$*, the probability mass function is given by*

   $$f_{Y|\{d_Y\}} = e^{-1.0684537582577533+0.7640712614003418y-0.12706332034025078y^2}, \quad y \in \{y_1, y_2, \ldots, y_9\}.$$

2. $\{y_1 = -20, -16, -12, -8, -4, 0, 4, 8, 12, 16, 20 = y_{11}\}$ *(here,* $N = 11$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 16.5741$*, the probability mass function is given by*

$$f_{Y|\{d_Y\}} = e^{-1.3377442247303293+0.5645861247447859y-0.08613916616052139y^2},$$
$$y \in \{y_1, y_2, \ldots, y_{11}\}.$$

3. $\{y_1 = -20, -18, -16, \ldots, 16, 18, 20 = y_{21}\}$ *(here,* $N = 21$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 16.5741$*, the probability mass function is given by*

$$f_{Y|\{d_Y\}} = e^{-2.0416173670052853+0.5721739130434783y-0.08695652173913045y^2},$$
$$y \in \{y_1, y_2, \ldots, y_{21}\}.$$

and thus, it is evidently clear that for $N = 21$*,* $\lambda_i^{(D)}$ *are as good as equal to* $\lambda_i^{(C)}$ *for* $i \in \{1, 2\}$*, together with* $\lambda_0^{(C)} + \log\left(\frac{20-(-20)}{21-1}\right) = -2.734764547565833 + \log(2) = -2.0416173670058875 \approx -2.0416173670052853 = \lambda_0^{(D)}$.

Example 3.3.2. *Here, we take* $a = -20$*,* $b = 20$*,* $m = 2$*,* $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 235.824$*.*

In accordance with the above data, the continuous approximated density function is given by

$$f_{Y|\{d_Y\}} = e^{-4.884382768986512+0.01423159215923913y+0.006393082832198655y^2},$$
$$-20 \leq y \leq 20.$$

So, corresponding to the discrete support

1. $\{y_1 = -20, -15, -10, -5, 0, 5, 10, 15, 20 = y_9\}$ *(here,* $N = 9$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 235.824$*, the probability mass function is given by*

$$f_{Y|\{d_Y\}} = e^{-2.80460516166883+0.01426494094022057y+0.002924326285777068y^2},$$
$$y \in \{y_1, y_2, \ldots, y_9\}.$$

2. $\{y_1 = -20, -16, -12, -8, -4, 0, 4, 8, 12, 16, 20 = y_{11}\}$ *(here,* $N = 11$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 235.824$*, the probability mass function is given by*

   $$f_{Y|\{d_Y\}} = e^{-3.090436143898487+0.014259657999991682y+0.0034088485135887973y^2},$$
   $$y \in \{y_1, y_2, \ldots, y_{11}\}.$$

3. $\{y_1 = -20, -18, -16, \ldots, 16, 18, 20 = y_{21}\}$ *(here,* $N = 21$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 235.824$*, the probability mass function is given by*

   $$f_{Y|\{d_Y\}} = e^{-3.9441237993721243+0.014247348475374011y+0.004621656553159281y^2},$$
   $$y \in \{y_1, y_2, \ldots, y_{21}\}.$$

4. $\{y_1 = -20, -19.5, -19, \ldots, 19, 19.5, 20 = y_{81}\}$ *(here,* $N = 81$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 235.824$*, the probability mass function is given by*

   $$f_{Y|\{d_Y\}} = e^{-5.503197448004097+0.014235948442220536y+0.005872522698977092y^2},$$
   $$y \in \{y_1, y_2, \ldots, y_{81}\}.$$

5. $\{y_1 = -20, -19.75, -19.5, \ldots, 19.5, 19.75, 20 = y_{161}\}$ *(here,* $N = 161$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 235.824$*, the probability mass function is given by*

   $$f_{Y|\{d_Y\}} = e^{-6.232165394628131+0.014233809818243154y+0.006124663159506933y^2},$$
   $$y \in \{y_1, y_2, \ldots, y_{161}\}.$$

6. $\{y_1 = -20, -19.875, -19.75, \ldots, 19.75, 19.875, 20 = y_{321}\}$ *(here,* $N = 321$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 235.824$*, the probability mass function is given by*

   $$f_{Y|\{d_Y\}} = e^{-6.944210924671172+0.014232711146443755y+0.006256722226090568y^2},$$
   $$y \in \{y_1, y_2, \ldots, y_{321}\}.$$

7. $\{y_1 = -20, -19.9375, -19.875 \ldots, 19.875, 19.9375, 20 = y_{641}\}$ *(here,* $N = 641$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 235.824$*, the probability mass function is given by*

   $$f_{Y|\{d_Y\}} = e^{-7.647072575058029+0.014232154227746513y+0.006324349319136263y^2},$$
   $$y \in \{y_1, y_2, \ldots, y_{641}\}.$$

8. $\{y_1 = -20, -19.9688, -19.9375 \ldots, 19.9375, 19.9688, 20 = y_{1281}\}$ *(here,* $N = 1281$*)*

   and with subject to $\mu_Y^{(1)} = 3.29$ *and* $\mu_Y^{(2)} = 235.824$*, the probability mass function is given by*

   $$f_{Y|\{d_Y\}} = e^{-8.345145788890813+0.014231873841696413y+0.006358575763627603y^2},$$
   $$y \in \{y_1, y_2, \ldots, y_{1281}\}.$$

and thus, it is evidently clear that for $N = 1281$*,* $\lambda_i^{(D)}$ *are well close to* $\lambda_i^{(C)}$ *for* $i \in \{1, 2\}$*, together with* $\lambda_0^{(C)} + \log\left(\frac{20-(-20)}{1281-1}\right) = -4.884382768986512 - \log(32) = -8.35011867178624 \approx -8.345145788890813 = \lambda_0^{(D)}$.

3.3.6 The existing classical moment problems

Let us state the existing moment problems with reference to [18] as well as to the page 2 of [59] by taking the random variable Y principally for continuous.

These moment problems, each of which deals with a **sequence of moments** of Y for the purpose of **determining the probability distribution** of Y, are hereby described as:

- the **Hamburger moment problem**, where the probability distribution of Y is supported by $(-\infty, +\infty)$
- the **Stieltjes moment problem**, where the probability distribution of Y is supported by $[0, +\infty)$
- the **Hausdorff moment problem**, where the probability distribution of Y is supported by $[a, b]$ $(a, b \in \mathbb{R})$. **Without any loss of generality**, the support of the probability distribution of Y can be taken for $[0, 1]$. This particular moment problem is subdivided to two subproblems, namely the **finite moment problem** (referred to a **finite sequence** of moments of Y) and the **infinite moment problem** (referred to a **infinite sequence** of moments of Y).

Though the **Hamburger moment problem** and the **Stieltjes moment problem** do not concern this dissertation in any way, we shall briefly state that the solvability of the Hamburger moment problem and Stieltjes moment problem with regard to the availability of a **finite** sequence of moments of Y has been discussed intensively in the paper [15]. This is briefly to say, that the solvability in this regard is theoretically possible.

The **Hausdorff's finite moment problem** with regard to the availability of a **finite** sequence of moments of Y is **correlated** to this dissertation (especially, a finite sequence of **two** moments for **continuous** cases of Y [8]). The **Hausdorff's infinite moment problem** is totally uncorrelated to this dissertation, except a kind of its simplicity is worth mentioning.

[8] the **discrete** cases of Y have a difficulty

3.4 The consistent density estimator

3.4.1 The background

Let us take the reference of the page 1 of [14], which says the following:

The idea of maximum entropy is simply to choose the probability density, which maximizes a particular entropy-measure with subject to a given set of moment constraints. The two desirable features of this methodology are

1. *existence of such a maximum entropy density given by first m assigned moments (the uniqueness of the maximum entropy density needs to be guaranteed).*

2. *the maximum entropy density should converge to the unknown density as the number of given moments increases.*

Notably,

1. The first feature precisely says that

   - the **probability mass function** $f_{Y|\{d_Y\}}(y_j) = e^{\lambda_0^{(D)} + \sum_{i=1}^{m} \lambda_i y_j^i}$, $j \in \{1, 2, \ldots, N\}$ of the **discrete** random variable Y is **uniquely** determinable by its first m moments $\mu_Y^{(1,D)}, \mu_Y^{(2,D)}, \ldots, \mu_Y^{(m,D)}$
   - the **probability density function** $f_{Y|\{d_Y\}}(y) = e^{\lambda_0^{(C)} + \sum_{i=1}^{m} \lambda_i y^i}$, $a \leq y \leq b$ of the **continuous** random variable Y is **uniquely** determinable by its first m moments $\mu_Y^{(1,C)}, \mu_Y^{(2,C)}, \ldots, \mu_Y^{(m,C)}$

   This particular feature shall be discussed in due course.

2. The second feature, which has been duly stated in [14], shall be intensively focussed in this section. To the best of my understanding, this feature has not been established so far.

Our **objective** in this section is to prove that the **probability density function** of the exponential polynomial distribution happens to be the **consistent density estimator** of the situation oriented need based unknown probability density function $f_Y(y)$, $a \leq y \leq b$. The **formal statement** of this **consistency character** shall be given in due course.

In due course, it shall be shown that the exponential polynomial probability distribution of the random variable Y (the polynomial being of a finite degree m, $m \in \mathbf{N}$) can be determined **uniquely** by the first predetermined m moments of the probability distribution of Y. This applies both in cases of discrete and continuous Y.

We have already seen that the probability density function $f_Y(y)$, $y \in [a,b]$ of Y is well approximately representable by an exponential polynomial density function denoted by $e^{p_Y(y)}$, $y \in [a,b]$, referred to (2.18), established by the **proposition 2.4.2**. In fact, (2.18) tells us that $e^{p_Y(y)}$, $a \leq y \leq b$ approximates $f_Y(y)$, $a \leq y \leq b$ good enough. Basically, we are to discuss the very issue that this approximation can **always** be bettered by increasing the value of m, **with regard** to the fact, that the first m moments of both the probability densities $f_Y(y)$, $a \leq y \leq b$ and $e^{p_Y(y)}$, $a \leq y \leq b$ are **identically the same**.

The aforesaid **consistency** is ascribed to the very fact that the goodness of this approximation can be **increased arbitrarily by increasing** m **arbitrarily**.

The derivation (3.95) belonging to the **proposition 3.3.3** (*with regard to the very fact that the moment* $\mu_Y^{(i,D)}$ *approaches the moment* $\mu_Y^{(i,C)}$ *in the* ***limiting sense*** *for every* $i \in \{1,2,\ldots,m\}$*, when* N *is made to be arbitrarily large*) **did not** assume that m increases proportionately with N, for eg. $m = N-1$, N being taken for the number of elements of $\mathcal{X}_Y$ in case Y is taken for discrete.

That is, **in contrary** to the **statement 3.3.1** our present discussion involves the case, when the number of elements of $\mathcal{X}_Y$ in case Y is taken for discrete is $N-1$ and $m = N-2$ (i.e. our present discussion is **independent** of the manner, in which we have deduced the relation (3.95). That is, we shall **ignore** the **statement 3.3.1** here). Precisely, in our present discussion in this **section 3.4**, m **does** increase proportionately with the increase in N.

3.4.2 The first lemma for the consistency

This lemma of this subsection described by the **subsequent proposition 3.4.1** necessitates the targeted usage of the **first mean value theorem for integrals**, which is well stated and proved in the page 169 of [17]. The statement of this very theorem is hereby given as follows:

Theorem 3.4.1 (First mean value theorem for integrals). *Let $f_Y(y)$ and $\phi_Y(y)$ be two Riemann integrable bounded functions in the closed interval $[c,d]$. Moreover, let $\phi_Y(y)$ keep the same sign in $[c,d]$. Then, there exists a real number k_Y, such that*

$$\int_c^d f_Y(y)\,\phi_Y(y)\,dy = k_Y \int_c^d \phi_Y(y)\,dy \tag{3.103}$$

where $l_Y \leq k_Y \leq u_Y$ with $l_Y = \inf_{y\in[c,d]} f_Y(y)$, $u_Y = \sup_{y\in[c,d]} f_Y(y)$

With this, we arrive at the following proposition:

Proposition 3.4.1 (Targeted application of the first mean value theorem for integrals 3.4.1). *Let the support of the probability density function $f_Y(y)$ of the **continuous** random variable Y be the closed interval $[a,b]$ as usual.*

*Now, if the interval $[a,b]$ is subdivided into a finite number of subintervals and if Y is taken for a **discrete** random variable, then the probability mass function of Y can be expressible as an exponential polynomial function.*

*Proof of the **proposition 3.4.1**.* In our case, since $f_Y(y)$ is continuous in $[a,b]$ (and understandably bounded in $[a,b]$), by setting

- $\delta = \frac{b-a}{N-1}$
- for any fixedly chosen j, $j \in \{1,2,\ldots,N-1\}$,
  - $c = a + (j-1)\delta$
  - $d = a + j\delta$
- $\phi_Y(y) = 1$, $c \leq y \leq d$

it is evidently clear that there exists at least one point $\mathbf{y}_j \in [a+(j-1)\delta, a+j\delta]$, such that $f_Y(\mathbf{y}_j) = k_Y$.

Therefore, for every $j \in \{1,2,\ldots,N-1\}$, by using the **theorem 3.4.1** stated by (3.103), we arrive at

$$\int_{a+(j-1)\delta}^{a+j\delta} f_Y(y)dy = f_Y(\mathbf{y}_j)\,(a + j\delta - (a + (j-1)\delta)) = f_Y(\mathbf{y}_j)\delta \tag{3.104}$$

which give the **individual probability elements** $f_Y(\mathbf{y}_j)\delta = \int\limits_{a+(j-1)\delta}^{a+j\delta} f_Y(y)dy$ ranging from $j = 1$ to $j = N-1$, but with subject to $\sum\limits_{j=1}^{N-1} f_Y(\mathbf{y}_j)\delta = 1$.

Therefore, by the **exact representation** of the probability mass function $f_Y(\mathbf{y}_j)\delta$, $j \in \{1, 2, \ldots, N-1\}$ referred to (2.9) established by the **proposition 2.4.1**, we get

$$f_Y(\mathbf{y}_j)\delta = f_Y(\mathbf{y}_j)\left(\frac{b-a}{N-1}\right) = e^{\lambda_0^{(D)}+\lambda_1\mathbf{y}_j+\lambda_2\mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}} \tag{3.105}$$

and this completes the proof of the **proposition 3.4.1**. □

3.4.3 The second lemma for the consistency

The lemma of this subsection describes the targeted usage of the stated **theorem 3.3.2**. This theorem however does not specifically say, whether the support of the probability distribution of Y is **closed** or **bounded**.

We shall write a **corollary** of the theorem 3.3.2 as a special case for our use, where the probability distribution of Y has a **compact support** as follows:

Corollary 3.4.1 (The special case of the theorem 3.3.2 of Shannon's and differential entropy). *If the probability density function of the continuous random variable Y denoted by $f_Y(y)$ is Riemann integrable, then*

$$\textbf{\textit{Shannon's entropy}} \text{ of } f_Y(\mathbf{y}_j)\delta + \log(\delta) \xrightarrow{\delta\to 0} \textbf{\textit{Differential entropy}} \text{ of } f_Y(y) \tag{3.106}$$

*where **in our case**, since the distributional support $\mathcal{X}_Y$ of Y is **compact**, we shall put the following to **usages**:*

- *δ = constant subinterval length belonging to the distributional support*
- *$f_Y(\mathbf{y}_j)\delta$ is the probability mass function of Y, if Y is discrete*
- *$f_Y(y)$ is the probability density function of Y, if Y is continuous*

*and in **contrary** to the **statement 3.3.1** (as we have already mentioned), we shall use $m = N-2$.*

With this, we shall proceed.

3.4.4 The consistency character

Referring to the page 11 of the paper [14], this consistency has been conjectured strongly. We shall go for the proof of the said consistency step by step. Let us state this consistency character formally at first.

Proposition 3.4.2 (Consistency of the density estimator). *The probability density function of Y defined by $e^{\lambda_0^{(C)}+\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_N y^N}$, $a \leq y \leq b$, $N \in \mathbb{N}$ is the **consistent density estimator** of the **existing** but **unknown** probability density function of Y defined by $f_Y(y)$, $a \leq y \leq b$.*

*Proof of the **proposition 3.4.2**.* With subject to the consideration of the discrete probability distribution given by (3.105) (of the **proposition 3.4.1** as an usage of the mean value theorem (3.103)), namely

$$\begin{aligned} f_Y(\mathbf{y}_j)\delta &= e^{\lambda_0^{(D)}+\lambda_1\mathbf{y}_j+\lambda_2\mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}} \\ \Leftrightarrow f_Y(\mathbf{y}_j) &= e^{\lambda_0^{(D)}-\log(\delta)+\lambda_1\mathbf{y}_j+\lambda_2\mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}}, j \in \{1,2,\ldots,N-1\} \end{aligned} \tag{3.107}$$

if δ is made infinitesimally smaller, then the following are to be carefully noted:

- $\mathbf{y}_j$ gets infinitesimally closer to y_j from **right** and to y_{j+1} from **left** for every $j \in \{1,2,\ldots,N-1\}$. In particular,
  - $\mathbf{y}_1$ gets infinitesimally closer to $a = y_1$ from **right**
  - $\mathbf{y}_{N-1}$ gets infinitesimally closer to $y_N = b$ from **left**
- the number of λ_j values, i.e. for $j \in \{1,2,\ldots,N-2\}$, namely $N-2$, gets arbitrarily larger. This happens because of $\delta = \frac{b-a}{N-1} \overset{N\to\infty}{\longrightarrow} 0$.
- each of the probability elements $p_j^{(\delta)} = e^{\lambda_0^{(D)}-\log(\delta)+\lambda_1\mathbf{y}_j+\lambda_2\mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}}$, $j \in \{1,2,\ldots,N-1\}$ gets arbitrarily smaller individually, but with subject to $\sum_{j=1}^{N-1} p_j^{(\delta)} = 1$.

Now, with regard to the consideration of (3.107), if $f_Y(y)$, $\mathbf{y}_1 \leq y \leq \mathbf{y}_{N-1}$ is taken for a **probability density function** of the continuous random

variable Y, we shall have to write

$$\begin{aligned} f_Y(y) &= e^{\lambda_0^{(C)}+\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_{N-2}y^{N-2}},\ \mathbf{y}_1 \leq y \leq \mathbf{y}_{N-1} \\ \text{where } e^{-\lambda_0^{(C)}} &= \int_{\mathbf{y}_1}^{\mathbf{y}_{N-1}} e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_{N-2}y^{N-2}}dy \end{aligned} \tag{3.108}$$

Again, with regard to

1. **infinitesimal smallness** of $\delta = \frac{b-a}{N-1}$
2. **continuity** of $f_Y(y)$, $a \leq y \leq b$ throughout the interval $[a, b]$

for any **arbitrarily small** ϵ, $\epsilon > 0$, there exists a $N_1 \in \mathbb{N}$, such that

$$\left| e^{\lambda_0^{(D)}-\log(\delta)+\lambda_1 \mathbf{y}_j+\lambda_2 \mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}} - e^{\lambda_0^{(C)}+\lambda_1 \mathbf{y}_j+\lambda_2 \mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}} \right| < \epsilon \tag{3.109}$$

for every $N > N_1$ and for every $j \in \{1, 2, \ldots, N-1\}$.

This principally says the following:

- $$\lambda_0^{(D)} - \log(\delta) \overset{\delta \to 0}{\longrightarrow} \lambda_0^{(C)} \tag{3.110}$$

  or equivalently

  $$\lambda_0^{(D)} - \log(\delta) \overset{N \to \infty}{\longrightarrow} \lambda_0^{(C)} \tag{3.111}$$

  and this happens to be affirmative to the well established (3.100)

- $$\begin{aligned} e^{\lambda_0^{(C)}+\lambda_1 \mathbf{y}_j+\lambda_2 \mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}} &\overset{\delta \to 0}{\longrightarrow} f_Y(\mathbf{y}_j) \\ &= e^{\lambda_0^{(D)}-\log(\delta)+\lambda_1 \mathbf{y}_j+\lambda_2 \mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}} \\ &\text{for every } j \in \{1, 2, \ldots, N-1\} \end{aligned} \tag{3.112}$$

  or equivalently

  $$\begin{aligned} e^{\lambda_0^{(C)}+\lambda_1 \mathbf{y}_j+\lambda_2 \mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}} &\overset{N \to \infty}{\longrightarrow} f_Y(\mathbf{y}_j) \\ &= e^{\lambda_0^{(D)}-\log(\delta)+\lambda_1 \mathbf{y}_j+\lambda_2 \mathbf{y}_j^2+\ldots+\lambda_{N-2}\mathbf{y}_j^{N-2}} \\ &\text{for every } j \in \{1, 2, \ldots, N-1\} \end{aligned} \tag{3.113}$$

Again, with regard to the very established statements $\mathbf{y}_j \overset{\delta\to 0}{\longrightarrow} y_j+$ and $\mathbf{y}_j \overset{\delta\to 0}{\longrightarrow} y_{j+1}-$ for every $j \in \{1, 2, \ldots, N-1\}$ especially $\mathbf{y}_1 \overset{\delta\to 0}{\longrightarrow} a+$ and $\mathbf{y}_{N-1} \overset{\delta\to 0}{\longrightarrow} b-$, together with the relation (3.108) giving the definition of $f_Y(y)$, $\mathbf{y}_1 \leq y \leq \mathbf{y}_{N-1}$ as a probability density function of the continuous Y, both (3.112) and (3.113) can be remodified to the following

$$e^{\lambda_0^{(C)}+\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_{N-2}y^{N-2}} \overset{\delta\to 0}{\longrightarrow} f_Y(y) \text{ for every } y \in [a, b] \tag{3.114}$$

or equivalently

$$e^{\lambda_0^{(C)}+\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_{N-2}y^{N-2}} \overset{N\to\infty}{\longrightarrow} f_Y(y) \text{ for every } y \in [a, b] \tag{3.115}$$

Exactly at this point, we shall make use of the **corollary 3.4.1** and in fact,

1. The **Shannon's entropy** of the probability distribution of the discrete Y defined by the probability mass function $f_Y(\mathbf{y}_j)\delta$, $j \in \{1, 2, \ldots, N-1\}$ is given by the expression $-\lambda_0^{(D)} - \lambda_1\mu_Y^{(1,D)} - \lambda_2\mu_Y^{(2,D)} - \ldots - \lambda_{N-2}\mu_Y^{(N-2,D)}$, $\mu_Y^{(i,D)}$ for $i \in \{1, 2, \ldots, N-2\}$ being the first $N-2$ moments of the **discrete** probability distribution.

2. The **differential entropy** of the probability distribution of the continuous Y defined by the probability density function $f_Y(y)$, $y \in [a, b]$ is given by the expression $-\lambda_0^{(C)} - \lambda_1\mu_Y^{(1,C)} - \lambda_2\mu_Y^{(2,C)} - \ldots - \lambda_{N-2}\mu_Y^{(N-2,C)}$, $\mu_Y^{(i,C)}$ for $i \in \{1, 2, \ldots, N-2\}$ being the first $N-2$ moments of the **continuous** probability distribution.

So, by the **corollary 3.4.1**, we get

$$\begin{aligned} &-\lambda_0^{(D)} - \lambda_1\mu_Y^{(1,D)} - \lambda_2\mu_Y^{(2,D)} - \ldots - \lambda_{N-2}\mu_Y^{(N-2,D)} + \log(\delta) \\ &\overset{\delta\to 0}{\longrightarrow} -\lambda_0^{(C)} - \lambda_1\mu_Y^{(1,C)} - \lambda_2\mu_Y^{(2,C)} - \ldots - \lambda_{N-2}\mu_Y^{(N-2,C)} \end{aligned} \tag{3.116}$$

and with subject to the statement (3.110) (or equivalently of (3.111)), namely $\lambda_0^{(D)} - \log(\delta) \overset{\delta\to 0}{\longrightarrow} \lambda_0^{(C)}$, (3.116) can be further simplified as

$$\begin{aligned} &\lambda_1\mu_Y^{(1,D)} + \lambda_2\mu_Y^{(2,D)} + \ldots + \lambda_{N-2}\mu_Y^{(N-2,D)} \\ &\overset{\delta\to 0}{\longrightarrow} \lambda_1\mu_Y^{(1,C)} + \lambda_2\mu_Y^{(2,C)} + \ldots + \lambda_{N-2}\mu_Y^{(N-2,C)} \end{aligned} \tag{3.117}$$

Therefore, by (3.117), in the **limiting sense**, the moments $\mu_Y^{(i,D)}$ are individually the same as the moments $\mu_Y^{(i,C)}$, $i \in \{1, 2, \ldots, N-2\}$. **Notably**, we

were **not** allowed to use the **corollary 3.3.3**, simply because this corollary refers to the **statement 3.3.1** that **contradicts** the assumption $m = N-2$.

Hence, by (3.115), we conclude that

$$e^{\lambda_0^{(C)}+\lambda_1 y+\lambda_2 y^2+...+\lambda_{N-2}y^{N-2}} \overset{N\to\infty}{\longrightarrow} f_Y(y) \text{ for every } y \in [a,b]$$

such that $\lambda_1, \lambda_2, \ldots, \lambda_{N-2}$ are **uniquely determinable** by the moments $\mu_Y^{(1,C)}, \mu_Y^{(2,C)}, \ldots, \mu_Y^{(N-2,C)}$, as already stated before.

This proves the **desired** consistency of the probability density estimator $e^{\lambda_0^{(C)}+\lambda_1 y+\lambda_2 y^2+...+\lambda_{N-2}y^{N-2}}$, $y \in [a,b]$ of the **unknown situation oriented need based** probability density function $f_Y(y)$, $y \in [a,b]$ of the **continuous random variable** Y. This proves the **proposition 3.4.2**. □

Note:

Remark 3.4.1 (The consistency character). *We need to state that the* ***consistency character*** *of the exponential polynomial distribution in continuous cases of Y duly elaborated in this section (the section 3.4) as well as the entropy convergence character stated in (3.118) (of the* ***remark 3.4.3****) are* ***purely theoretical derivations*** *meant for establishing the principle characteristic properties of the exponential polynomial distribution.*

From the ***practical or rather from the stochastic point of view****, only the first two moments of Y have basic practical significance.*

Remark 3.4.2. *The realizations of this* ***consistency character*** *is duly demonstrated in form of* ***graphical representations*** *in the coming chapter 8 of* ***Illustrations of M.I. Probability Distributions***

Remark 3.4.3 (Convergence in entropy). *An immediate consequence of the (as established)* ***consistency of the probability density estimator*** *is the* ***entropy convergence*** *or in other words, the* ***convergence in entropy****. The entropy convergence (defined in the page 8 of [41]) says that, if $H(f_{Y|\{d_Y^{(n)}\}}(y))$ is the differential entropy of the probability density $f_{Y|\{d_Y^{(n)}\}}(y) = e^{\lambda_0^{(C)}+\lambda_1 y+\lambda_2 y^2+...+\lambda_n y^n}$, $a \le y \le b$, such that $d_Y^{(n)} = \left(\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(n)}\right)$, then the sequence defined by $\left\{H(f_{Y|\{d_Y^{(n)}\}}(y))\right\}_{n\in\mathbb{N}}$ converges to the limit $H(f_Y(y))$, the limit being the differential entropy of the*

unknown but existing probability density $f_Y(y)$ *having identically the same moments* $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(n)}$.

In fact, the following has been established (in the paper [41])

$$\begin{aligned} &H(f_{Y|\{d_Y^{(0)}\}}(y)) \geq H(f_{Y|\{d_Y^{(1)}\}}(y)) \geq H(f_{Y|\{d_Y^{(2)}\}}(y)) \geq \ldots \\ &\geq H(f_{Y|\{d_Y^{(n)}\}}(y)) \geq \ldots \geq H(f_Y(y)) \end{aligned} \tag{3.118}$$

Remark 3.4.4. *In cases of Stieltjes and Hamburger's moment problems (referred to the page 8 of [15]), as in the above case, the sequence of maximum entropy approximating densities converge* ***in entropy*** *to the density characterized by its moments. This convergence problem has also been treated widely in the Hausdorff's moment problem in [9].*

3.5 The minimum information selection principle

There are many methods developed in probability theory and statistics for selecting a probability distribution in a given situation. A majority of them uses highly adaptable probability measures which are fitted to the observed data. Without aiming at the completeness, below a list of frequently applied methods is given.

- Combinatorial methods based on equiprobability.
- Asymptotic methods based on limits.
- Empirical methods based on data.
- Fitting methods based on universal distributions.
- Maximum Entropy methods based on the available knowledge.

Note, that none of these methods is explicitly based on the results of a scientific investigation of the process under consideration. Apart from these above stated methods, our method for selecting a suitable probability distribution is based on the **minimum information principle**, which targets exclusively the scientific need and the observation of the empirical value of d_Y. In a plain language, our method is based on the scientific investigation of the process under consideration.

In the following course of discussions, an alternative method to the traditional methods, which has been proposed in [54], is outlined. It starts with identifying minimum requirements with respect to the necessary information for describing a situation appropriately.

For determining the **minimum amount of information** needed to select an appropriate probability measure $P_{Y|\{d_Y\}}$ in case of a discrete Y (or an appropriate probability density $f_{Y|\{d_Y\}}$, in case of a continuous Y), it is assumed that enough experience is made or enough knowledge is available about the process in question and the random variable of interest Y for deciding the question about the co-domain of the random structure function $\mathcal{P}$. If this necessary knowledge is not available, a meaningful description of the random structure of Y is impossible and no attempts should be made to describe the

same, but must be postponed until enough experience has been gathered. Thus, the following function $\mathcal{P}$ (as discussed) has to be structured as:

$$\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_m \tag{3.119}$$

With subject to the description (3.119) of $\mathcal{P}$, as we have already mentioned, the domain of $\mathcal{P}$, namely $\mathcal{T}_{D_Y}(\mathcal{D}_Y)$ is assumed (**for the sake of simplicity only**) to contain the singletons denoted by $\{d_Y\}$ only, i.e. $d_Y \in \mathcal{D}_Y$. In the section 4.3 of the chapter 4, it has been shown that each d_Y determines an element of $\mathbb{P}_m$ (i.e. a probability distribution) uniquely.

We shall however confine our detailed discussions to $m \leq 2$. In other words, the co-domain of $\mathcal{P}$, namely $\mathbb{P}_m$ is either the **constant family**, the **monotone family** or the **uni-extremal family**.

3.5.1 Constant family: $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_0$

If $\mathcal{P}(\{d_Y\}) \in \mathbb{P}_0$, then the minimum amount of information necessary for describing the situation coincides with the complete knowledge, which is given by the range of variability $\mathcal{X}_Y(\{d_Y\})$ of the random variable $Y|\{d_Y\}$.

$$\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_0 \tag{3.120}$$

$$\mathcal{P}(\{d_Y\}) = f_{Y|\{d_Y\}} \text{ with} \tag{3.121}$$

$$f_{Y|\{d_Y\}}(y) = \frac{1}{|\mathcal{X}_Y(\{d_Y\})|} \text{ for } y \in \mathcal{X}_Y(\{d_Y\}) \tag{3.122}$$

3.5.2 Monotonic family: $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_1$

If $\mathcal{P}(\{d_Y\}) \in \mathbb{P}_1$ holds, it is known, that the probability mass or density function is a **monotonic decreasing** or a **monotonic increasing** function, and the same holds for the series $\sum_{i=0}^{\infty} \lambda_i(d_Y) y^i$. Thus, the question arises, how many terms of the series $\sum_{i=0}^{\infty} \lambda_i(d_Y) y^i$ are necessary for describing the monotone random structure in the given situation.

The above question cannot be answered without having further information about the situation. However, the question of the **minimum** number of

terms necessarily needed to meet the property of monotonicity can be easily answered. The simplest monotonic function is a linear function of y, i.e.

$$\lambda_0(d_Y) + \lambda_1(d_Y)y \text{ with } \lambda_1(d_Y) \neq 0 \tag{3.123}$$

Consequently, (3.123) is selected to be the power of e, yielding a random structure function $\mathcal{P}(\{d_Y\})$ with the following probability mass or probability density function:

$$f_{Y|\{d_Y\}}(y) = e^{\lambda_0(d_Y)+\lambda_1(d_Y)y} \text{ for } y \in \mathcal{X}_Y(\{d_Y\}) \tag{3.124}$$

In a monotone situation, let the search for an appropriate element of $\mathbb{P}_1$ be restricted to the probability mass or probability density function of the form (3.124). This search demands a minimum amount of information necessary to determine the coefficients $\lambda_0(d_Y)$ and $\lambda_1(d_Y)$. Therefore, the selection criterium applied above is called *minimum information principle*.

Of course, in addition to the monotonic property, if other properties of the probability mass or probability density function are known, these properties should be taken into account.

3.5.3 Uni-extremal family: $\mathcal{P} : \mathcal{T}_{D_Y}(\mathcal{D}_Y) \rightarrow \mathbb{P}_2$

If $\mathcal{P}(\{d_Y\}) \in \mathbb{P}_2$ holds, it is known, that the probability mass or probability density function has exactly one relative extremum, implying that the series $\sum_{i=0}^{\infty} \lambda_i(d_Y)y^i$ too has exactly one relative extremum. Thus, in a similar manner as in the previous case, the minimum information principle yields a quadratic function

$$\lambda_0(d_Y) + \lambda_1(d_Y)y + \lambda_2(d_Y)y^2 \text{ with } \lambda_2(d_Y) \neq 0 \tag{3.125}$$

as the power of e. Therefore, the random structure function $\mathcal{P}(\{d_Y\})$ is selected accordingly and thereby yielding the following probability mass or probability density function:

$$f_{Y|\{d_Y\}}(y) = e^{\lambda_0(d_Y)+\lambda_1(d_Y)y+\lambda_2(d_Y)y^2} \text{ for } y \in \mathcal{X}_Y(\{d_Y\}) \tag{3.126}$$

with $\lambda_2(d_Y) \neq 0$.

3.5.4 The basis of the computation of λ_i, $i \in \{1, 2, \ldots, m\}$ values

The computation of λ_i, $i \in \{1, 2, \ldots, m\}$ values in my dissertation is **based** exclusively on the **minimum information principle**. This means that the probability distribution of the **desired type** (i.e. the type being defined by the **number of extremes**, namely $m - 1$ number of extremes in this case) shall exclusively be the **starting point** for the computation of $\lambda_1, \lambda_2, \ldots, \lambda_m$.

In other words, the user of my software programs has to say specifically, which **probability distribution type** he wishes.

This is precisely to say that the amount of available quantitative information given by the **number of moments** shall **not** be the starting point of the same. This amount shall be put under consideration, **only** after the **type** of the desired probability distribution is specifically known.

Notably, for stochastic procedures, $m \leq 2$. This means, the multi-extremal probability distributions, under normal circumstances, are **not** asked for.

3.6 Conclusive remarks

3.6.1 The monotone and uni-extremal probability distributions

Understandably, the monotone and the uni-extremal probability distributions given by (3.124) and (3.126) respectively are of simplest forms. Now, let us consider the following probability density functions:

$$f_{Y|\{d_Y\}}(y) = e^{-1.3024282635577977+2y+\frac{1}{3}y^3}, \ 0 \leq y \leq 1 \tag{3.127}$$

$$f_{Y|\{d_Y\}}(y) = e^{0.07357162236102023-\frac{1}{2}y+\frac{1}{2}y^2-\frac{1}{6}y^3+\frac{1}{4}y^4}, \ 0 \leq y \leq 1 \tag{3.128}$$

and represent them graphically as follows:

The graphical representation of the probability density function (3.127) given by the figure (3.1) clearly shows a continuous **monotonic character** within its support $0 \leq y \leq 1$. This is an example of a probability distribution belonging to the family $\mathbb{P}_1$, though this is **not** a continuous monotonic minimum information probability distribution. By minimum information principle, only the first moment would have been necessary as the needed information

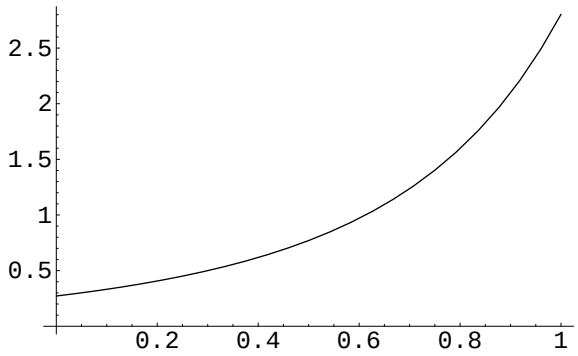


Figure 3.1: The Probability Density $f_{Y|\{d_Y\}}(y)$, $0 \leq y \leq 1$ referred to (3.127)

for the construction of this probability distribution, apart from the predetermined support $[0, 1]$. In fact, the first three moments have been needed as the information here.

The graphical representation of the probability density function (3.128) given by the figure (3.2) clearly shows a continuous **uni-extremal character** within its support $0 \leq y \leq 1$. This is an example of a probability distribution belonging to the family $\mathbb{P}_2$, though this **not** a continuous uni-extremal minimum information probability distribution. By minimum information principle, only the first and the second moment would have been necessary as the needed information for the construction of this probability distribution, apart from the predetermined support $[0, 1]$. In fact, the first four moments have been needed as the information here.

Conclusively, in each of the two above illustrated examples, bigger amount of numerical work as well as longer program running time was needed for the construction of the probability distributions. This is precisely to conclude that the usage of the minimum information principle basically **minimizes** the **amount of numerical work** for solving the equations involving moments as well as the **amount of program running time**.

Importantly, it has to be mentioned that, if the probability distribution of Y needs to be uni- extremal, then the second moment of Y is ought not to be within certain bounds, otherwise the desired **uni- extremal character**

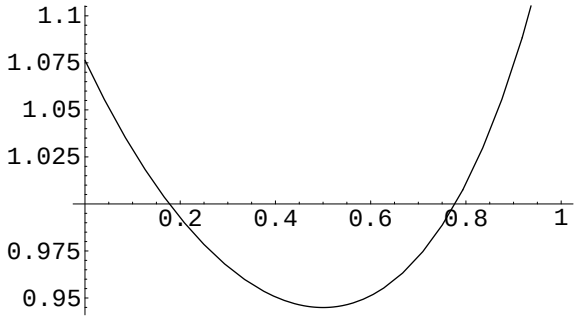


Figure 3.2: The Probability Density $f_{Y|\{d_Y\}}(y)$, $0 \leq y \leq 1$ referred to (3.128)

will be violated. This particular point has been intensively elaborated in the subsection 6.3.15 **principally for continuous cases**. The discussions about the **discrete cases** are largely similar.

3.6.2 A general important remark

Each additional relative extreme of a probability mass or density function signals a more **complicated situation** demanding additional information.

In **practice**, the exact values of moments are never available, but **estimated values** of the same. Therefore, if additional information (namely additional moments) are used to construct a probability distribution, then there is an every possibility that the desired type of the probability distribution is hopelessly violated, because the number of extremes (of the probability distribution) is **different** from the same of the desired type. This difference is solely due to errors in estimations.

From this point of view, the maximum entropy principle speaks about the **theory**, but the minimum information principle speaks about the **practice**.

3.7 The simplicity in the Hausdorff's infinite moment problem

In case the readers of this dissertation ask about the case, if infinite number of moments of a random variable, say X, may be taken into consideration, we shall consider this case briefly. We shall also include few important comments in this regard.

This very consideration may be a part of an abstract mathematical analysis, but does have any part in the science of stochastics, simply because the **concept of infinity** is purely a mathematical concept and does not have any significance in the real world, as far as **realistic sizes** are concerned. Therefore, this discussion has no relevance in this dissertation.

The mathematician Felix Hausdorff (1868 - 1942) established the infinite moment problem. The statement of this theorem is given as follows:

Theorem 3.7.1 (Hausdorff's infinite moment problem). *Corresponding to a sequence of moments given by* $\{\mu_m\}_{m\in\mathbb{N}_0}$ *there exists a distribution function* $\Psi(x)$, $0 \leq x \leq 1$, *such that*

$$\mu_m = \int_{x=0}^{x=1} x^m d\Psi(x) \tag{3.129}$$

if and only if the sequence is completely monotonic, where $\{\mu_m\}_{m\in\mathbb{N}_0}$ *is completely monotonic means*

$$(-1)^k \boldsymbol{\Delta}^k \mu_m \geq 0, \quad k \in \mathbb{N}_0 \tag{3.130}$$

The **theorem 3.7.1** is referred to the Hausdorff's theorem 1 given in the page 193 of [29].

Definition 3.7.1 (Brief introduction of the $\boldsymbol{\Delta}$ symbol).

$$\begin{aligned} \boldsymbol{\Delta}^0 \mu_m &= \mu_m, \quad \boldsymbol{\Delta}^1 \mu_m = \mu_{m+1} - \mu_m \\ \boldsymbol{\Delta}^k \mu_m &= \boldsymbol{\Delta}^{k-1} \mu_{m+1} - \boldsymbol{\Delta}^{k-1} \mu_m, \ \ k \geq 1 \\ &= \sum_{p=0}^{k} (-1)^p \binom{k}{p} \mu_{m+k-p} \end{aligned} \tag{3.131}$$

3.7.1 The simplicity of Hausdorff's moment problem

With subject to the consideration of the discrete case of X, the Hausdorff's theorem 1 (i.e. the theorem 3.7.1) primarily says about the **bounds of the moments** as far as the existence of a probability distribution of X is concerned. It **does not** however say that the existence of the (discrete) probability distribution of X is possible for every **predetermined** range of variability $\mathcal{X}_X = \{0 = x_1, x_2, \ldots, x_N = 1\}$. It only says that if the sequence of moments of X, namely if the sequence $\{\mu_m\}_{m\in\mathbb{N}_0}$ is completely monotonic (Hausdorff's condition (3.130)) then there exists a $\mathcal{X}_X$, such that a **discrete** probability distribution of X exists.

Basically, the theorem 3.7.1 states the necessary and the sufficient condition of the existence for the **continuous** probability distribution of X.

Now, let us prove the following proposition:

Proposition 3.7.1. *The **theorem 3.7.1** defined by the Hausdorff's conditional statement (3.130) is **completely deducible** with the help of the following underlying condition (3.132) named **as the condition of complete monotonicity** addressed to **all the moments of** X, i.e. the **moments** μ_n, such that*

$$\mu_{n+1} < \mu_n \text{ for } \textbf{every } n \in \mathbb{N}_0 \tag{3.132}$$

*In other words, the (Hausdorff's) **theorem 3.7.1** defined by the conditional statement (3.130) is **implied by** the (aforesaid) condition (3.132).*

*Proof of the **proposition 3.7.1**.* In accordance with the condition (3.132) as well as by the definition of the difference operator $\boldsymbol{\Delta}$, i.e. $\boldsymbol{\Delta}\mu_n = \mu_{n+1} - \mu_n$, $n \in \mathbb{N}_0$, we arrive at

$$\begin{aligned} -\boldsymbol{\Delta}\mu_n &= \mu_n - \mu_{n+1} = E[X^n(1-X)] > 0 \\ -\boldsymbol{\Delta}\mu_{n+1} &= \mu_{n+1} - \mu_{n+2} = E[X^{n+1}(1-X)] > 0 \end{aligned}$$

from which, we can easily get

$$-\boldsymbol{\Delta}\mu_n - (-\boldsymbol{\Delta}\mu_{n+1}) = -\boldsymbol{\Delta}(\mu_n - \mu_{n+1}) = (-\boldsymbol{\Delta})^2\mu_n = E[X^n(1-X)^2] > 0$$

such that $(-\boldsymbol{\Delta})^2\mu_n$ means that the operator $-\boldsymbol{\Delta}$ is operated on $(-\boldsymbol{\Delta})\mu_n$.

Again, by using the very rule $(-\boldsymbol{\Delta})^2\mu_n = E[X^n(1-X)^2]$, $n \in \mathbb{N}_0$ we get

$$\begin{aligned} (-\boldsymbol{\Delta})^2\mu_n &= E[X^n(1-X)^2] > 0 \\ (-\boldsymbol{\Delta})^2\mu_{n+1} &= E[X^{n+1}(1-X)^2] > 0 \end{aligned}$$

from which, exactly in the similar manner, we can easily get

$$(-\boldsymbol{\Delta})^2\mu_n - ((-\boldsymbol{\Delta})^2\mu_{n+1}) = (-\boldsymbol{\Delta})^2(\mu_n - \mu_{n+1}) = (-\boldsymbol{\Delta})^3\mu_n$$
$$= E[X^n(1-X)^2] - E[X^{n+1}(1-X)^2] = E[X^n(1-X)^3] > 0$$

which basically concludes that $(-\boldsymbol{\Delta})^3\mu_n = E[X^n(1-X)^3] > 0$.

Proceeding exactly in this manner, for any arbitrarily chosen $k \in \mathbb{N}_0$, we finally arrive at

$$(-\boldsymbol{\Delta})^k\mu_n = (-1)^k\boldsymbol{\Delta}^k\mu_n = E[X^n(1-X)^k] > 0, \text{ which gives (3.130)}$$

Hence, the **desired proof** of the **proposition 3.7.1**. □

Remark 3.7.1. *This aforesaid condition (3.132) is rather obvious and therefore is purely trivial and this* ***triviality*** *is precisely the* ***simplicity of the Hausdorff's infinite moment problem****. In other words, this trivial condition (3.132) is* ***enough*** *to describe the condition for the existence of the probability distribution of X.*

Remark 3.7.2. *Clearly, the* ***violation*** *of condition (3.132) for* ***any*** *value of n, $n \in \mathbb{N}_0$ makes sure that the existence of the probability distribution of X is* ***completely ruled out****.*

As a brief remark, I would like to address my readers to take a note of the following statement: My first former supervisor misguided me to use this Hausdorff's statement (3.130) for finding out the bounds of the moments of X for my dissertational work, though probably unintentionally. Only at a later point of time I could find out that this very Hausdorff's statement is not relevant for my work.

Chapter 4

General m. i. probability distributions

4.1 Introductory statements

In this chapter we shall discuss about the nature of the minimum information probability distributions (with compact supports).

Before we go ahead, we state that, purely for the sake of **simplicity**, we shall use the symbols such as $\mathcal{D}_Y$ (symbolizing the ignorance space with regard to Y); Y shall be **mostly** used instead of $Y|\{d_Y\}$, X instead of $X|\{d\}$; λ_i instead of $\lambda_i(d_Y)$ in the coming course of discussions. Moreover the frequently used symbols $\mathcal{X}_Y$, $\mathcal{X}_X$, a and b stand purely for $\mathcal{X}_Y(\{d_Y\})$, $\mathcal{X}_X(\{d\})$, $a(d_Y)$ and $b(d_Y)$ **respectively**, **unless** any clarity in this regard is of utmost importance.

Each of the moments of a probability distribution has a bounded range, i.e. bounded both **above** and **below**, the details of these lower and the upper bounds of each of such moments shall be discussed in due course. **Importantly**, the **entire course** of this chapter assumes that **each** of these moments lies between its **greatest lower bound** and its **least upper bound**.

Therefore, we restate that the **approximating** probability distribution is constructed by means of the following:

- the given compact support
- the available moments of the random variable

With this, in the light of the **approximating efficiency** of the exponential polynomial probability distribution discussed in the **section 2.4**, let us give the following statements describing the **approximating** (for continuous cases) or **exact** (for discrete cases) probability distribution of a random variable Y:

Statement 4.1.1. *If Y is a discrete random variable, whose range of variability is $\mathcal{X}_Y(\{d_Y\}) = \{a = y_1, y_2, \dots y_N = b\}$, then its **approximating** or **exact** probability mass function $f_{Y|\{d_Y\}}$ is given by*

$$f_{Y|\{d_Y\}}(y_j) = e^{\lambda_0+\lambda_1 y_j+\lambda_2 y_j^2+\dots+\lambda_m y_j^m}, \quad j = 1, 2, \dots, N$$

*such that $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)}, \dots, \mu_Y^{(m)})$. Here, the polynomial $\sum_{i=0}^{m} \lambda_i y_j^i$ in y_j can have the **highest allowable** power of y_j to be $N-1$, **because** the range of variability of Y (i.e. the support of the probability distribution of Y) consists of exactly N elements.*

*That is, $m \leq N-1$ **must** hold, otherwise the **unique** determination of $\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_m$ is **not possible** (where m stands for the number of available moments of Y) and this very fact we shall prove formally subsequently. **Notably**, if $m = N-1$, then the probability distribution of Y defined by the probability mass function $f_{Y|\{d_Y\}}$, as we have already discussed, becomes **exact**.*

Statement 4.1.2. *If Y is a continuous random variable, whose range of variability is $\mathcal{X}_Y(\{d_Y\}) = [a, b]$, then its **approximating** probability density function $f_{Y|\{d_Y\}}$ is given by*

$$f_{Y|\{d_Y\}}(y) = e^{\lambda_0+\lambda_1 y+\lambda_2 y^2+\dots+\lambda_m y^m}$$

such that $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)}, \dots, \mu_Y^{(m)})$.

In each of the above cases, the probability distribution of Y has, as we have already discussed, the **maximum entropy**.

Our primarily next objective shall be to show that the moments of Y, namely $1 = \mu_Y^{(0)}, \mu_Y^{(1)}, \mu_Y^{(2)}, \dots, \mu_Y^{(m)}$, $(m \in \mathbb{N})$ determine its probability distribution (both in discrete and continuous cases), **uniquely**.

4.2 The systems of simultaneous equations involving moments

With subject to the given range of variability $\mathcal{X}_Y(\{d_Y\})$ and the deterministic variable $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)})$,
we shall proceed to prove that the probability distribution of Y (i.e. probability mass function in case Y is discrete or probability density function in case Y is continuous), described by

$$f_{Y|\{d_Y\}}(y) = e^{\lambda_0+\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}, \;\; y \in \mathcal{X}_Y(\{d_Y\}) \tag{4.1}$$

can be determined uniquely, i.e. the solution of the following system of equations in $\lambda_1, \lambda_2, \ldots, \lambda_m$ is unique:

$$\begin{cases} \mu_Y^{(1)} & = & \frac{\int\limits_{\mathcal{X}_Y} y\, e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)}{\int\limits_{\mathcal{X}_Y} e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)} \\ \mu_Y^{(2)} & = & \frac{\int\limits_{\mathcal{X}_Y} y^2\, e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)}{\int\limits_{\mathcal{X}_Y} e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)} \\ \quad\vdots & & \\ \mu_Y^{(m)} & = & \frac{\int\limits_{\mathcal{X}_Y} y^m\, e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)}{\int\limits_{\mathcal{X}_Y} e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)} \end{cases} \tag{4.2}$$

Note, that the above **integral notation** stands for a **finite summation** if Y is a **discrete** random variable or for a **Riemann integral** within the finite limits a and b if Y is a **continuous** random variable.

In order to discuss the **uniqueness** of the solution of (4.2), we **conveniently** need to introduce another random variable $X|\{d\}$, defined by

$$X = X|\{d\} = \frac{Y|\{d_Y\} - a}{b-a} = \frac{Y-a}{b-a} \tag{4.3}$$

which necessarily means that the range of variability of $X|\{d\}$ is compact as well. It is symbolized and given by $\mathcal{X}_X(\{d\}) = \{0 = x_1, x_2, \ldots, x_N = 1\}$ (for

a discrete X) or $\mathcal{X}_X(\{d\}) = [0,1]$ (for a continuous X), is bounded by the limits 0 and 1.

The first m ($m \in \mathbb{N}$) moments of the transformed random variable X are the components of the deterministic variable d, viz.

$$d = (\mu_1, \mu_2, \ldots, \mu_m)$$

where $1 = \mu_0, \mu_1, \mu_2, \ldots, \mu_m$ are the corresponding moments of X up to the order of m.

It is clear that the uniqueness of the solution of (4.2) is tantamount to the uniqueness of the solution of the following system (4.4) of equations in $\beta_1, \beta_2, \ldots, \beta_m$:

$$\begin{cases} \mu_1 & = & \frac{\int\limits_{\mathcal{X}_X} x\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} \\ \mu_2 & = & \frac{\int\limits_{\mathcal{X}_X} x^2\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} \\ \vdots & & \\ \mu_m & = & \frac{\int\limits_{\mathcal{X}_X} x^m\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} \end{cases} \tag{4.4}$$

Note, that the above used integral notation, as usual, stands for a finite summation, if $X|\{d\}$ is a discrete random variable or for a bounded integral within the limits 0 and 1, if $X|\{d\}$ is a continuous random variable.

Therefore, the probability distribution of X is given by

$$f_{X|\{d\}}(x) = e^{\beta_0+\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m},\ \ x \in \mathcal{X}_X(\{d\}) \tag{4.5}$$

with subject to

$$\beta_0 = -\log\left(\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)\right) \tag{4.6}$$

Thus, from the **practical point of view**, the **computed solution** of (4.2) is **not yielded directly**, but by solving (4.4) for $(\beta_1, \beta_2, \ldots, \beta_m)$ at first and then finding the values of $\lambda_1, \lambda_2, \ldots, \lambda_m$ uniquely by comparing the coefficients of all the positive integral powers of y of both the sides of the following relation:

$$\lambda_1 y + \lambda_2 y^2 + \ldots + \lambda_m y^m = \beta_1 \left(\frac{y-a}{b-a}\right) + \beta_2 \left(\frac{y-a}{b-a}\right)^2 + \ldots + \beta_m \left(\frac{y-a}{b-a}\right)^m \tag{4.7}$$

Therefore, it is clearly evident, that the uniqueness of the solution of (4.4) is tantamount to the uniqueness of the solution of (4.2).

The uniqueness of the solution of (4.2) can be proven directly. **Equivalently**, the uniqueness of the solution of (4.4) can be proven to prove the uniqueness of (4.2). That is, **both are possible**. As a matter of fact, the uniqueness of the solution of (4.2) can be proven directly by means of the well known Gibbs' inequality [1] (i.e. without the usage of the system (4.4) of equations) in a **more simpler** way. So, before we go ahead to prove the same in the very next section, we state the Gibbs' inequality at first:

Proposition 4.2.1 (Gibbs' inequality). *If two probability distributions described by f_Y and g_Y have the identically common bounded support $\mathcal{X}_Y$, then the Gibbs' inequality states that the Kullback-Leibler divergence is always non-negative, i. e.*

$$\int_{\mathcal{X}_Y} f_Y(y) \log\left(\frac{f_Y(y)}{g_Y(y)}\right) \nu(dy) \geq 0 \tag{4.8}$$

Moreover, the equality in (4.8) holds according to the following rule:

$$\int_{\mathcal{X}_Y} f_Y(y) \log\left(\frac{f_Y(y)}{g_Y(y)}\right) \nu(dy) = 0 \Leftrightarrow f_Y(y) = g_Y(y) \ \ \nu - \textit{almost everywhere} \tag{4.9}$$

The proof of the **proposition 4.2.1** by means of Jensen's inequality is referred to the page 562 of [33].

[1] established by J. Willard Gibbs in the nineteenth century

4.3 Uniqueness of the solution of the equation-system for $m \in \mathbb{N}$

For every $m \in \mathbb{N}_0$, the solution of the system of simultaneous equations (4.2), **if exists**, is **unique**, provided $N \geq m+1$ in cases when Y happens to be discrete. This is the statement, which we shall establish in this very section by using (4.8) of the **Gibb's inequality** together with the consequential special case of equality in (4.8) stated by (4.9).

Notably, **alternatively**, this aforesaid uniqueness can also be established by **minimizing** the functional $\Gamma(\lambda_1, \lambda_2, \ldots, \lambda_m)$ defined by (3.89) globally. The statement of this very uniqueness are duly given by some authors (referred to the pages 2 to 3 of [41] as well as to the page 3 of [59]).

The formal statement is hereby given as follows:

Theorem 4.3.1 (Uniqueness of the solution of the simultaneous system of equations of moments). *The solution of the system of m simultaneous equations given in (4.2), namely*

$$\begin{cases} \mu_Y^{(1)} & = & \frac{\int\limits_{\mathcal{X}_Y} y\ e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)}{\int\limits_{\mathcal{X}_Y} e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)} \\ \mu_Y^{(2)} & = & \frac{\int\limits_{\mathcal{X}_Y} y^2\ e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)}{\int\limits_{\mathcal{X}_Y} e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)} \\ \quad\vdots & & \\ \mu_Y^{(m)} & = & \frac{\int\limits_{\mathcal{X}_Y} y^m\ e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)}{\int\limits_{\mathcal{X}_Y} e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)} \end{cases} \tag{4.2}$$

*is unique, provided $N \geq m+1$ holds in **discrete** cases.*

*Proof of the **theorem 4.3.1**.* For the sake of definiteness, corresponding to a fixedly given range of variability $\mathcal{X}_Y$ of the random variable Y, let Y have two probability distributions described by $f_{Y|\{d_Y\}}(y)$ and $g_{Y|\{d_Y\}}(y)$ having the identically same moments $\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)}$.

So, if we set

$$f_{Y|\{d_Y\}}(y) = e^{\lambda_0+\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m} \tag{4.10}$$

$$g_{Y|\{d_Y\}}(y) = e^{\lambda_0^{'}+\lambda_1^{'} y+\lambda_2^{'} y^2+\ldots+\lambda_m^{'} y^m} \tag{4.11}$$

such that

$$e^{\lambda_0} = \frac{1}{\int\limits_{\mathcal{X}_Y} e^{\lambda_1 y+\lambda_2 y^2+\ldots+\lambda_m y^m}\nu(dy)} \tag{4.12}$$

$$e^{\lambda_0^{'}} = \frac{1}{\int\limits_{\mathcal{X}_Y} e^{\lambda_1^{'} y+\lambda_2^{'} y^2+\ldots+\lambda_m^{'} y^m}\nu(dy)} \tag{4.13}$$

Therefore, according to the given condition, we have

$$\mu_Y^{(i)} = \int\limits_{\mathcal{X}_Y} y^i f_{Y|\{d_Y\}}(y)\nu(dy) = \int\limits_{\mathcal{X}_Y} y^i g_{Y|\{d_Y\}}(y)\nu(dy),\ i = 0, 1, 2, \ldots, m \tag{4.14}$$

where $\mu_Y^{(0)} = 1$.

Now, by taking the relative entropy of $f_{Y|\{d_Y\}}$ with respect to $g_{Y|\{d_Y\}}$, in accordance with the (4.8) of the **Gibbs' inequality**, we get

$$\begin{aligned}\int\limits_{\mathcal{X}_Y} f_{Y|\{d_Y\}}(y)\log\left(\frac{f_{Y|\{d_Y\}}(y)}{g_{Y|\{d_Y\}}(y)}\right)\nu(dy) &= \int\limits_{\mathcal{X}_Y} f_{Y|\{d_Y\}}(y)\sum_{i=0}^{m}\left(\lambda_i-\lambda_i^{'}\right)y^i\nu(dy)\\ &= \sum_{i=0}^{m}\left(\lambda_i-\lambda_i^{'}\right)\mu_Y^{(i)}\\ &\geq 0\end{aligned} \tag{4.15}$$

Again, by taking the relative entropy of $g_{Y|\{d_Y\}}$ with respect to $f_{Y|\{d_Y\}}$, in accordance with the (4.8) of the **Gibbs' inequality**, exactly in the same way, we get

$$\begin{aligned}\int\limits_{\mathcal{X}_Y} g_{Y|\{d_Y\}}(y)\log\left(\frac{g_{Y|\{d_Y\}}(y)}{f_{Y|\{d_Y\}}(y)}\right)\nu(dy) &= \int\limits_{\mathcal{X}_Y} g_{Y|\{d_Y\}}(y)\sum_{i=0}^{m}\left(\lambda_i^{'}-\lambda_i\right)y^i\nu(dy)\\ &= \sum_{i=0}^{m}\left(\lambda_i^{'}-\lambda_i\right)\mu_Y^{(i)}\\ &\geq 0\end{aligned} \tag{4.16}$$

From (4.15) and (4.16) we can solely conclude that

$$\sum_{i=0}^{m}\left(\lambda_i-\lambda_i^{'}\right)\mu_Y^{(i)}=\sum_{i=0}^{m}\left(\lambda_i^{'}-\lambda_i\right)\mu_Y^{(i)}=0 \tag{4.17}$$

which means nothing, but the relative entropy of $f_{Y|\{d_Y\}}$ with respect to $g_{Y|\{d_Y\}}$ and/or vice versa is zero. Hence, by the (4.9) of the **Gibbs' equality**,

$$f_{Y|\{d_Y\}}(y)=g_{Y|\{d_Y\}}(y)\ \ \nu-\ \text{almost everywhere} \tag{4.18}$$

This basically implies that

$$\begin{aligned}&\sum_{i=0}^{m}\left(\lambda_i-\lambda_i^{'}\right)y^i=0\ \ \nu-\ \text{almost everywhere}\\ &\Rightarrow\lambda_i=\lambda_i^{'}\ \text{for}\ i=0,1,2,\ldots,m\end{aligned} \tag{4.19}$$

by the fundamental theorem of Algebra.

Hence, the two probability distributions of Y, namely $f_{Y|\{d_Y\}}$ and $g_{Y|\{d_Y\}}$ are not distinct and therefore the solution of the system (4.2) is unique. This completes the proof of the **theorem 4.3.1**. □

Subsequently, our next step shall be to **discuss the (unique) existence of the solution of the system (4.2)** by means of certain Hankel matrices.

4.4 Existence of the solution of the equation-system for $m \in \mathbb{N}$

Obviously, the **existence of the solution of (4.4) is tantamount to the existence of the solution of (4.2)**. Our principle aim shall be to state the necessary and the sufficient condition for the existence of the solution of (4.2). This shall therefore be fulfilled by confining ourselves to the system of equations (4.4) only, **without any loss of generality**.

Since the dissertation primarily concerns the developments and the character analysis of **probability distributions** of **standard types** only, the **formal rigorous proofs** of the existence of the solution of (4.4) for $m=1$ and $m=2$ are given in subsections 6.2.3 and 6.3.10. Otherwise, the formal statements

for the existence of the solution of (4.4) for $m \geq 3$ are given without a rigorous proof.

Again, since the solution of the solution of the system of equations (4.2) has already been shown to be **unique with subject to its existence**, the solution of the system of equations (4.4) is **equivalently unique as well with subject to its existence**.

In other words, with regard to the existence of the unique solution, both the systems of equations (4.2) and (4.4) have the same characteristic property.

Keeping this in mind, let us discuss the **Hausdorff's finite moment problem** as the immediate next course of our action.

4.4.1 Hausdorff's finite moment problem

The Hausdorff's finite moment problem (**defined** in the page 2 of [60]) is precisely the problem of the **determination** of a probability density function $\mathbf{f}_X(x)$, $0 \leq x \leq 1$ with subject to the given moment constraints, namely

$$\mu_i = \int_0^1 x^i \mathbf{f}_X(x) dx, \text{ for } i \in \{0, 1, 2, \ldots, m\} \tag{4.20}$$

Even a brief discussion of the Hausdorff's finite moment problem (the problem may be appropriately abbreviated as **HFMP**) necessitates the introduction of the moment space. The moment space (defined in the page 3 of [41]) is a set of points, each of which has m components. The definition of the moment space reads as follows:

Definition 4.4.1 (Moment space). *Let* $\mathbf{D}^m \subset \mathbb{R}_+^m$ *be the* ***convex hull*** *of the* ***closed and bounded set*** $\{(x, x^2, \ldots, x^m) | 0 \leq x \leq 1\}$. *Then,* $\mathbf{D}^m$ *is called the* m- ***moment space*** *and is a convex set.*

In that case, the following **three definitional propositions**, which are duly elaborated in [20], are hereby given as:

Definition and Proposition 4.4.1 (No solution exists). *If the point* $\overrightarrow{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$ *is outside* $\mathbf{D}^m$, *then the* ***HFMP does not*** *admit any solution.*

Definition and Proposition 4.4.2 (A solution exists uniquely on the boundary). *If the point* $\overrightarrow{\mu} = (\mu_1, \mu_2, \ldots, \mu_m) \in \partial\mathbf{D}^m$ *(*$\partial\mathbf{D}^m$ *is the boundary of* $\mathbf{D}^m$*), then the* ***HFMP*** *admits* ***only*** *one probability distribution* [2] *having* $\mu_1, \mu_2, \ldots, \mu_m$ *as the first* m *moments. In this regard, this* ***unique*** *probability distribution is* ***not of continuous type*** *represented by the density function* $\mathbf{f}_X(x)$, $0 \leq x \leq 1$, *but a* ***discrete*** *probability distribution with a finite support.*

Definition and Proposition 4.4.3 (Infinitely many solutions exist). *If the point* $\overrightarrow{\mu} = (\mu_1, \mu_2, \ldots, \mu_m) \in int\mathbf{D}^m$ *(*$int\mathbf{D}^m$ *is the interior of* $\mathbf{D}^m$*), then the* ***HFMP*** *admits* ***infinitely many*** *probability distributions.*

4.4.2 The necessary and sufficient condition

Theorem 4.4.1 (Hausdorff's necessary and sufficient condition for the finite moment problem). *The necessary and sufficient condition for the existence of* ***a*** *solution to the* ***HFMP****, namely a solution of the system of equations (4.20) is stated equivalently in the two following ways:*

- *the point* $\overrightarrow{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$ *is an* ***interior point*** *of the moment space* $\mathbf{D}^m$ *(as stated in the* ***definition and proposition 4.4.3*** *)*
- *By taking* $i \in \{1, 2, \ldots, m\}$*, such that* $i \in \{2n, 2n+1\}$ *according as* ***i*** *is* ***even*** *or* ***odd*** *with regard to* $2n \leq m$ *as well as* $2n+1 \leq m$*, each of the Hankel matrices* [3]*, namely* $\underline{H}_{2n}$ *(defined by (****4.23****)),* $\underline{H}_{2n+1}$ *(defined by (****4.27****)),* $\overline{H}_{2n}$ *(defined by (****4.30****)) and* $\overline{H}_{2n+1}$ *(defined by (****4.33****)) is individually* ***positive definite****.*

The statement of the **theorem 4.4.1** is referred to the theorem 3.1 given in the page 7 of [41] as well as to the page 2 of [60]. The proof of the **theorem 4.4.1** is referred to [20]. The statement of the theorem 4.4.1 is shall be elaborately explained, but a formal general proof shall **not** be given.

Nextly, these aforesaid Hankel matrices $\underline{H}_{2n}$, $\underline{H}_{2n+1}$, $\overline{H}_{2n}$ and $\overline{H}_{2n+1}$ (with subject to $n \in \mathbb{N}$, but $2n \leq m$, $2n+1 \leq m$), whose positive definiteness is the **necessary and sufficient condition** for the existence of a solution of the system of equations (4.20), shall be **elaborately defined** (the indexation of moments of X contained in these Hankel matrices is in accordance with the **statement 2.4.2**).

[2] The probability distribution in this regard is representable **probabilistically**

[3] Definitions of Hankel matrices are referred to the page 3 of [41] & to the page 2 of [60]

Firstly, we state the Hankel matrices, whose **positive** determinants give the **exact greatest lower bounds** of the moments μ_i, $i \in \{1, 2, \ldots, m\}$:

Definition 4.4.2 (Hankel matrix for the determination of the greatest lower bound of μ_i for an even i).

$$\underline{H}_2 = \begin{pmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \tag{4.21}$$

$$\underline{H}_4 = \begin{pmatrix} 1 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{pmatrix} \tag{4.22}$$

By proceeding exactly in this way, for any $i = 2n$, such that $n \in \{1, 2, \ldots, \lfloor \frac{m}{2} \rfloor\}$, we arrive at

$$\underline{H}_{2n} = \begin{pmatrix} 1 & \mu_1 & \mu_2 & \ldots & \mu_n \\ \mu_1 & \mu_2 & \mu_3 & \ldots & \mu_{n+1} \\ \mu_2 & \mu_3 & \mu_4 & \ldots & \mu_{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_n & \mu_{n+1} & \mu_{n+2} & \ldots & \mu_{2n} \end{pmatrix} \tag{4.23}$$

Definition 4.4.3 (Hankel matrix for the determination of the greatest lower bound of μ_i for an odd i).

$$\underline{H}_1 = \mu_1 \tag{4.24}$$

$$\underline{H}_3 = \begin{pmatrix} \mu_1 & \mu_2 \\ \mu_2 & \mu_3 \end{pmatrix} \tag{4.25}$$

$$\underline{H}_5 = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \\ \mu_3 & \mu_4 & \mu_5 \end{pmatrix} \tag{4.26}$$

By proceeding exactly in this way, for any $i = 2n + 1$, such that $n \in \{0, 1, 2, \ldots, \lfloor \frac{m-1}{2} \rfloor\}$,we arrive at

$$\underline{H}_{2n+1} = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{n+1} \\ \mu_2 & \mu_3 & \mu_4 & \cdots & \mu_{n+2} \\ \mu_3 & \mu_4 & \mu_5 & \cdots & \mu_{n+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_{n+1} & \mu_{n+2} & \mu_{n+3} & \cdots & \mu_{2n+1} \end{pmatrix} \tag{4.27}$$

Secondly, we state the Hankel matrices, whose determinants give the **exact least upper bounds** of the moments μ_i, $i \in \{1, 2, \ldots, m\}$:

Definition 4.4.4 (Hankel matrix for the determination of the least upper bound of μ_i for an even i).

$$\overline{H}_2 = \mu_1 - \mu_2 \tag{4.28}$$

$$\overline{H}_4 = \begin{pmatrix} \mu_1 - \mu_2 & \mu_2 - \mu_3 \\ \mu_2 - \mu_3 & \mu_3 - \mu_4 \end{pmatrix} \tag{4.29}$$

By proceeding exactly in this way, for any $i = 2n$, such that $n \in \{1, 2, \ldots, \lfloor \frac{m}{2} \rfloor\}$, we arrive at

$$\overline{H}_{2n} = \begin{pmatrix} \mu_1 - \mu_2 & \mu_2 - \mu_3 & \mu_3 - \mu_4 & \ldots & \mu_n - \mu_{n+1} \\ \mu_2 - \mu_3 & \mu_3 - \mu_4 & \mu_4 - \mu_5 & \ldots & \mu_{n+1} - \mu_{n+2} \\ \mu_3 - \mu_4 & \mu_4 - \mu_5 & \mu_5 - \mu_6 & \ldots & \mu_{n+2} - \mu_{n+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_n - \mu_{n+1} & \mu_{n+1} - \mu_{n+2} & \mu_{n+2} - \mu_{n+3} & \ldots & \mu_{2n-1} - \mu_{2n} \end{pmatrix} \tag{4.30}$$

Definition 4.4.5 (Hankel matrix for the determination of the least upper bound of μ_i for an odd i).

$$\overline{H}_1 = 1 - \mu_1 \tag{4.31}$$

$$\overline{H}_3 = \begin{pmatrix} 1-\mu_1 & \mu_1-\mu_2 \\ \mu_1-\mu_2 & \mu_2-\mu_3 \end{pmatrix} \tag{4.32}$$

By proceeding exactly in this way, for any $i = 2n+1$, such that $n \in \{0, 1, 2, \ldots, \lfloor \frac{m-1}{2} \rfloor\}$, we arrive at

$$\overline{H}_{2n+1} = \begin{pmatrix} 1-\mu_1 & \mu_1-\mu_2 & \mu_2-\mu_3 & \cdots & \mu_n-\mu_{n+1} \\ \mu_1-\mu_2 & \mu_2-\mu_3 & \mu_3-\mu_4 & \cdots & \mu_{n+1}-\mu_{n+2} \\ \mu_2-\mu_3 & \mu_3-\mu_4 & \mu_4-\mu_5 & \cdots & \mu_{n+2}-\mu_{n+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_n-\mu_{n+1} & \mu_{n+1}-\mu_{n+2} & \mu_{n+2}-\mu_{n+3} & \cdots & \mu_{2n}-\mu_{2n+1} \end{pmatrix} \tag{4.33}$$

The elaborate explanation of the statement of the necessary and sufficient condition of the existence of the solution of the system (4.20) (i.e. the formal description of the statement of the **theorem 4.4.1**) for any finitely fixed $m \in \mathbb{N}$ is thereby complete. As we have already mentioned, the formal rigorous proof of the **theorem 4.4.1** for any $m \in \mathbb{N}$ shall not be given in this dissertation.

However, the formal rigorous proof of the **theorem 4.4.1** in the **individual cases** for $m \in \{1, 2\}$, which are **just relevant for this dissertation**, are duly given in the subsections (6.2.3) and (6.3.10). In course of proving the same, both the **geometrical** and the **analytical** side of the problems are **intensively discussed** and are of extreme importance for the **character analysis of standard minimum information probability distributions**.

Remark 4.4.1. *As a brief note, the name addressed to the above stated* ***theorem 4.4.1*** *can be* ***precisely*** *put as the* ***Hausdorff's condition*** *for the finite moment problem.*

This Hausdorff's condition is basically addressed to the ***continuous case*** *of the random variable X only. If X happens to be discrete, this Hausdorff's condition needs to be handled a bit* ***differently****.*

Remark 4.4.2. *The* ***Hausdorff's finite moment problem****, in general, refers to probability distributions of* ***any*** *kind. The exponential polynomial probability distribution is only* ***one of such kinds*** *of probability distributions, which* ***HFMP*** *refers to.*

Remark 4.4.3 (Contextual realization of the Hausdorff's finite moment problem). *Let us take the exponential polynomial probability density function $f_{X|\{d^{(n)}\}}(x) = e^{\beta_0+\beta_1 x+\beta_2 x^2+\ldots+\beta_n x^n}$, $0 \leq x \leq 1$ determined by means of the first n moments, namely $\mu_1, \mu_2, \ldots, \mu_m, \mu_{m+1}, \ldots, \mu_n$, (i.e. $n \geq m$). Here, the* ***definition and proposition 4.4.3*** *basically says that, if the first m moments, namely $\mu_1, \mu_2, \ldots, \mu_m$ are* ***fixed****, then there are* ***infinitely many*** *such probability densities $f_{X|\{d^{(n)}\}}(x)$ when the rest $n - m$ moments, namely $\mu_{m+1}, \mu_{m+2}, \ldots, \mu_n$* ***vary****.*

Corollary 4.4.1 (The unique existence of the system of equations (4.4)). *By the* ***theorem 4.3.1****,* ***if a*** *solution of the system of equations (4.4),* ***at all exists****, then the solution is* ***unique****.*

Again, by the ***theorem 4.4.1****, the* ***necessary and sufficient condition*** *for the* ***existence*** *of* ***a*** *solution of the system of equations (4.4) is the positive definiteness of* ***each*** *of the Hankel matrices (4.23), (4.27), (4.30), and (4.33).*

Hence, conclusively, with subject to the above stated ***positive definiteness****, the probability density function* $f_{X|\{d\}}(x) = e^{\beta_0+\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}$, $0 \leq x \leq 1$ *(* $\beta_0 = -\log\left(\int\limits_0^1 e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m} dx\right)$ *) is* ***uniquely determinable****, which gets proved by the proof of the* ***unique existence*** *of* $\beta_1, \beta_2, \ldots, \beta_m$ *as the solution of the system of equations (4.4) with subject to the predetermined moments* $\mu_1, \mu_2, \ldots, \mu_m$ *of the random variable* X*.*

4.5 The necessary criterion for the solvability

4.5.1 The question addressed to the solvability

An extremely important question arises, whether the system of equations (4.4) is **at all generally** solvable by the **Newton-Raphson** method, **with subject to the fulfillment** of the **Newton-Raphson's convergence criterion** (duly stated and rigorously proved in the subsection 12.3.3) is **fulfilled**. The answer is **yes** and we need to establish this assertion formally.

This establishment of our assertion lies solely on the **positive definiteness** of a particular $m \times m$ square matrix.

The establishment of the solvability criterion necessitates a remodification of the system of equations (4.4) by setting

$$\overrightarrow{\boldsymbol{\beta}} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} \in \mathbb{R}^m \tag{4.34}$$

and subsequently by defining the function $\overrightarrow{\mathbf{f}}$ as

$$\overrightarrow{\mathbf{f}}(\overrightarrow{\boldsymbol{\beta}}) = \overrightarrow{\mathbf{f}} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} = \begin{pmatrix} f_1(\overrightarrow{\boldsymbol{\beta}}) \\ f_2(\overrightarrow{\boldsymbol{\beta}}) \\ \vdots \\ f_m(\overrightarrow{\boldsymbol{\beta}}) \end{pmatrix} = \overrightarrow{\mathbf{0}} \tag{4.35}$$

such that

$$\begin{cases} f_1(\overrightarrow{\boldsymbol{\beta}}) & = & \frac{\int\limits_{\mathcal{X}_X} x\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} - \mu_1^* = 0 \\ f_2(\overrightarrow{\boldsymbol{\beta}}) & = & \frac{\int\limits_{\mathcal{X}_X} x^2\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} - \mu_2^* = 0 \\ \vdots & & \\ f_m(\overrightarrow{\boldsymbol{\beta}}) & = & \frac{\int\limits_{\mathcal{X}_X} x^m\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} - \mu_m^* = 0 \end{cases} \tag{4.36}$$

where $\mu_1^*, \mu_2^*, \ldots, \mu_m^*$ are the user-given predetermined values of the moments of the random variable X and therefore are purely **constants** with respect to the variables $\beta_1, \beta_2, \ldots, \beta_m$. Here, we needed to use these new symbols, namely $\mu_1^*, \mu_2^*, \ldots, \mu_m^*$, because we simply need to differentiate between $\mu_i = \frac{\int\limits_{\mathcal{X}_X} x^i\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}$ and μ_i^* for $i \in \{1, 2, \ldots, m\}$ here.

However, this clear differentiation between μ_i and μ_i^* is not always necessary for every task. **If the meanings and interpretations are completely unambiguous, then we can easily write μ_i instead of μ_i^*,** $i \in \{1, 2, \ldots, m\}$. In fact, we shall do this in course of our coming discussions.

Clearly, apart from the consideration of this aforesaid differentiation, the systems of equations (4.4) and (4.36) are **exactly** the same. So, we shall focus our immediate present discussions on the solvability of the system of equations (4.36) **equivalently**.

Let the (**unique**) solution of (4.4) (or equivalently of (4.36)) be denoted by

$$\overrightarrow{\boldsymbol{\beta}}^* = \begin{pmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_m^* \end{pmatrix} \in \mathbb{R}^m \tag{4.37}$$

which means, $\overrightarrow{\mathbf{f}}(\overrightarrow{\boldsymbol{\beta}}^*) = \overrightarrow{\mathbf{0}}$.

Again, let the (initial) **starting** solution of (4.36) be denoted by

$$\overrightarrow{\boldsymbol{\beta}}_0 = \begin{pmatrix} \beta_1^{(0)} \\ \beta_2^{(0)} \\ \vdots \\ \beta_m^{(0)} \end{pmatrix} \in \mathbb{R}^m \tag{4.38}$$

which has to be feeded to the Newton-Raphson procedure, **provided** $\overrightarrow{\boldsymbol{\beta}}_0$ lies within the **convergence radius** of $\overrightarrow{\boldsymbol{\beta}}^*$, the intensive discussions of which have been given in the subsection 12.3.3.

Each procedural step of the Newton-Raphson procedure (referred to (12.57)) is restated as

$$\overrightarrow{\boldsymbol{\beta}}_{k+1} = \overrightarrow{\boldsymbol{\beta}}_k - \left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}}_k) \right)^{-1} \overrightarrow{\mathbf{f}}(\overrightarrow{\boldsymbol{\beta}}_k) \text{ for every } k \in \mathbb{N}_0$$

Exactly at this point, we are in a position to state that the solvability of the Newton-Raphson necessitates the **existence** of the quantity $\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}}_k) \right)^{-1}$ or in other words, the **invertibility** of $\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}}_k)$.

With this, before we could finally arrive at the formal statement of the solvability criterion, we need to state and prove two important lemmas. Before this, let us give the expression of $\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}})$ in the following way:

Definition and Proposition 4.5.1 ($\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}})$). *The derivative of* $\overrightarrow{\mathbf{f}}(\overrightarrow{\boldsymbol{\beta}})$ *for any* ***arbitrarily*** *chosen* $\overrightarrow{\boldsymbol{\beta}} \in \mathbb{R}^m$*, namely* $\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}})$ *is*

$$\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}}) = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_1}{\partial \beta_2} & \cdots & \frac{\partial \mu_1}{\partial \beta_m} \\ \frac{\partial \mu_2}{\partial \beta_1} & \frac{\partial \mu_2}{\partial \beta_2} & \cdots & \frac{\partial \mu_2}{\partial \beta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_m}{\partial \beta_1} & \frac{\partial \mu_m}{\partial \beta_2} & \cdots & \frac{\partial \mu_m}{\partial \beta_m} \end{pmatrix} \tag{4.39}$$

Proof. The **proof** follows immediately by **deriving** each μ_i with respect to β_j **partially** for $i, j \in \{1, 2, \ldots, m\}$ in accordance with the system of simultaneous equations (4.36). □

4.5.2 The first lemma addressed to a covariance matrix

Definition 4.5.1 (A special covariance matrix). *The square matrix, denoted by* $\mathsf{Cov}[\mathbf{X}_{(m)}]$*, is the covariance matrix of the* $\boldsymbol{m}$*- dimensional random vector defined by*

$$\mathbf{X}_{(m)} = \left(X, X^2, \ldots, X^m\right) \tag{4.40}$$

In this subsection, we shall show that $\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}})$ happens to be equal to the covariance matrix $\mathsf{Cov}[\mathbf{X}_{(m)}]$ of the random vector $\mathbf{X}_{(m)} = (X, X^2, \ldots, X^m)$.

Therefore, we shall state and prove the following proposition accordingly:

Proposition 4.5.1. $\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}}) = \mathsf{Cov}[\mathbf{X}_{(\boldsymbol{m})}]$ *for every* $\overrightarrow{\boldsymbol{\beta}} \in \mathbb{R}^m$

Proof of the ***proposition 4.5.1****.* Since the partial derivative $\frac{\partial \mu_i}{\partial \beta_j}$ is the element of the i^{th} row and j^{th} column of the matrix $\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}})$, we need to prove that it is also the same of that of the covariance matrix $\mathsf{Cov}[\mathbf{X}_{(m)}]$.

The element of the i^{th} row and j^{th} column of $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is the covariance of X^i and X^j denoted by $A_{i,j} = E\left[(X^i - \mu_i)(X^j - \mu_j)\right]$, where $1 \leq i, j \leq m$. This element $E\left[(X^i - \mu_i)(X^j - \mu_j)\right]$ shall also be denoted by $\sigma_{i,j}$, though the same is normally symbolized by $\mu_{i,j}$ (referred to the page 155 of [13]).

All what we need, is basically to prove the expression of $\sigma_{i,j}$ stated in (4.44), namely

$$\sigma_{i,j} = \frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_j}{\partial \beta_i} = \mu_{i+j} - \mu_i \mu_j, \quad \text{such that } i, j \in \{1, 2, \ldots, m\} \tag{4.44}$$

By the system (4.4), for every $i \in \{1, 2, \ldots, m\}$, we have

$$\mu_i = \frac{\int\limits_{\mathcal{X}_X} x^i \, e^{\beta_1 x + \beta_2 x^2 + \ldots + \beta_m x^m} \nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x + \beta_2 x^2 + \ldots + \beta_m x^m} \nu_X(dx)} \tag{4.41}$$

Now, the question arises, **whether** we can at all **interchange** the **operations** of **integration** $\int\limits_{\mathcal{X}_X}$ and the **partial differentiation** $\frac{\partial}{\partial \beta_j}$ in the process of differentiating μ_i partially with respect to β_j, for $i, j \in \{1, 2, \ldots, m\}$ with regard to the above relation (4.41) or not. The answer is **yes**. Now, let us

distinguish the two cases, namely when the random variable X is **discrete** and **continuous** one by one:

If X is **discrete**, then each of the two integrals $\int\limits_{\mathcal{X}_X} x^i \, e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m} \nu_X(dx)$ and $\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m} \nu_X(dx)$ having **continuous integrands** belonging to (4.41) is a **finite sum**. Here, the the integration operation $\int\limits_{\mathcal{X}_X}$ is a finite summation $\sum\limits_{j=1}^{N}$. Hence, the operations $\sum\limits_{j=1}^{N}$ and $\frac{\partial}{\partial \beta_j}$ are **interchangeable**.

Again, if X is **continuous**, then each of the two integrals $\int\limits_{\mathcal{X}_X} x^i \, e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m} \nu_X(dx)$ and $\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m} \nu_X(dx)$ having **continuous integrands** belonging to (4.41) is a **Riemann integral**. In this case, let us refer to the Theorem 2 titled by **Differentiation of the Integral** of page 306 of [17]. This theorem of *differentiation of the integral* states that

Let $\phi(y) = \int\limits_a^b f(x,y)dx$, *where* $f(x,y)$ *is a continuous function of* (x,y) *in the rectangle* $R : \{a \leq x \leq b, c \leq y \leq d\}$ *and* $f_y(x,y)$ *exists and is continuous in* R. *Then* $\phi^{'}(y) = \frac{d}{dy}\left\{\int\limits_a^b f(x,y)dx\right\}$ *exists and is equal to* $\int\limits_a^b \left(\frac{\partial}{\partial y} f(x,y)\right) dx$

In our case, the integrand $e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}$ is continuous for every β_j, $j \in \{1, 2, \ldots, m\}$ and $-\infty < \beta_j < \infty$. So, the theorem of **differentiation of the integral** is **well applicable** here. In other words, the operations $\int\limits_0^1$ and $\frac{\partial}{\partial \beta_j}$ are **interchangeable** here as well.

Thus, by keeping this **interchangeability** in **both discrete and continuous cases** of X in mind, by taking $p_X(x) = \beta_1 x + \beta_2 x^2 + \ldots + \beta_m x^m$, we have for every $j \in \{1, 2, \ldots, m\}$

$$\begin{aligned}
&\frac{\partial \mu_i}{\partial \beta_j} \\
&= \frac{\left(\int\limits_{\mathcal{X}_X} e^{p_X(x)}\nu_X(dx)\right)\left(\int\limits_{\mathcal{X}_X} x^{i+j}e^{p_X(x)}\nu_X(dx)\right)-\left(\int\limits_{\mathcal{X}_X} x^{i}e^{p_X(x)}\nu_X(dx)\right)\left(\int\limits_{\mathcal{X}_X} x^{j}e^{p_X(x)}\nu_X(dx)\right)}{\left(\int\limits_{x\in\mathcal{X}_X} e^{p_X(x)}\nu_X(dx)\right)^2} \\
&= \mu_{i+j} - \mu_i\mu_j = \sigma_{i,j} \qquad \square
\end{aligned} \tag{4.42}$$

and thereby proving our **proposition 4.5.1**. $\square$

4.5.3 The second lemma addressed to the positive definiteness of $\mathsf{Cov}[\mathbf{X}_{(m)}]$

We shall state and prove the following proposition in this subsection:

Proposition 4.5.2 (The positive definiteness). *For every $m \in \mathbb{N}$, the covariance matrix $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is positive definite, provided $N \geq m+1$ in cases, when X happens to be discrete.*

*Proof of the **proposition 4.5.2**.* The covariance (symmetric) matrix of $\mathbf{X}_{(m)} = (X, X^2, \ldots, X^m)$, being symbolized by $\mathsf{Cov}[\mathbf{X}_{(m)}]$, is given by

$$\mathsf{Cov}[\mathbf{X}_{(m)}] = \left(E\left[(X^i-\mu_i)(X^j-\mu_j)\right]\right)_{1\leq i\,,\,j\leq m} = \begin{pmatrix} A_{1,1} & A_{1,2} & \ldots & A_{1,m} \\ A_{2,1} & A_{2,2} & \ldots & A_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \ldots & A_{m,m} \end{pmatrix}$$

Obviously, $\mathsf{Cov}[\mathbf{X}_{(m)}]$ shall also be denoted by $(\sigma_{i,j})_{1\leq i\,,\,j\leq m}$ and is expressed as

$$(\sigma_{i,j})_{1\leq i\,,\,j\leq m} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_1}{\partial \beta_2} & \cdots & \frac{\partial \mu_1}{\partial \beta_m} \\ \frac{\partial \mu_2}{\partial \beta_1} & \frac{\partial \mu_2}{\partial \beta_2} & \cdots & \frac{\partial \mu_2}{\partial \beta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_m}{\partial \beta_1} & \frac{\partial \mu_m}{\partial \beta_2} & \cdots & \frac{\partial \mu_m}{\partial \beta_m} \end{pmatrix} \tag{4.43}$$

whose symmetry is well established by using (4.4) and by the very fact that

$$\sigma_{i,j} = \frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_j}{\partial \beta_i} = \mu_{i+j} - \mu_i \mu_j = \sigma_i \sigma_j \rho_{i,j}, \text{ such that } i, j \in \{1, 2, \ldots, m\}. \tag{4.44}$$

such that σ_i, σ_j, $\sigma_{i,j}$ and $\rho_{i,j}$ are the symbols of **variance** of X^i, **variance** of X^j, **covariance** of X^i and X^j and **correlation coefficient** of X^i and X^j respectively.

Therefore, keeping these things in mind, we give the proceed as follows:

The very fact that every covariance matrix is **positive semi-definite** has been well stated and **proven** in the page T.11.3, theorem number T.11.1.4 of [12].

Thus, for $m \in \mathbb{N}$, $\mathsf{Cov}[\mathbf{X}_{(m)}]$ could either be positive semi definite or positive definite.

Now, in order to prove the positive definiteness of $\mathsf{Cov}[\mathbf{X}_{(m)}]$, we necessarily need to make use of the following result (referred to the proven theorem T.11.1.5 in the page T.11.4 of [12]):

The $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is not positive definite, i.e. positive semi definite

$$\iff \tag{4.45}$$

$\exists$ scalars $\alpha_1, \alpha_2, \ldots, \alpha_m$ and $\widehat{\alpha}$ with respect to $\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_m^2 \neq 0$, i.e. not all α_is $(i = 1, 2, \ldots, m)$ are simultaneously zero, such that

$$P_{X|\{d\}}\left(\left\{x | \alpha_1 x + \alpha_2 x^2 + \ldots \alpha_m x^m = \widehat{\alpha}\right\}\right) = 1 \tag{4.46}$$

To go ahead with our operation by using the immediately stated above proven result, we shall treat the discrete and continuous cases separately:

Discrete Case: X is discrete:

In this case, the range of variability of X containing N elements is $\mathcal{X}_X(\{d\}) = \{0 = x_1, x_2, \ldots, x_N = 1\}$

Accordingly, with subject to the hypothetical assumption that $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is not positive definite, we would arrive at the following system of N equations

in $\alpha_1, \alpha_2, \ldots, \alpha_m$:

$$\begin{cases} \alpha_1 x_1 + \alpha_2 x_1^2 + \ldots + \alpha_m x_1^m & = \widehat{\alpha} \\ \alpha_1 x_2 + \alpha_2 x_2^2 + \ldots + \alpha_m x_2^m & = \widehat{\alpha} \\ \alpha_1 x_3 + \alpha_2 x_3^2 + \ldots + \alpha_m x_3^m & = \widehat{\alpha} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots & \vdots \\ \alpha_1 x_N + \alpha_2 x_N^2 + \ldots + \alpha_m x_N^m & = \widehat{\alpha} \end{cases} \tag{4.47}$$

Now, with subject to $x_1 = 0$, by the first equation of the above system of equations, we get

$$\alpha_1(0) + \alpha_2(0) + \ldots + \alpha_m(0) = \widehat{\alpha} \Rightarrow \widehat{\alpha} = 0 \tag{4.48}$$

and with subject to $x_N = 1$, by the last equation of the above system of equations, we get

$$\alpha_1 + \alpha_2 + \ldots + \alpha_m = \widehat{\alpha} = 0 \tag{4.49}$$

and thus the above system of N equations with m unknowns (4.47) basically reduces to the following system of $N-1$ equations with m unknowns:

$$\begin{cases} \alpha_1 + \alpha_2 + \ldots + \alpha_m & = 0 \\ \alpha_1 x_2 + \alpha_2 x_2^2 + \ldots + \alpha_m x_2^m & = 0 \\ \alpha_1 x_3 + \alpha_2 x_3^2 + \ldots + \alpha_m x_3^m & = 0 \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots & \vdots \\ \alpha_1 x_{N-1} + \alpha_2 x_{N-1}^2 + \ldots + \alpha_m x_{N-1}^m & = 0 \end{cases} \tag{4.50}$$

So, the question arises, whether the system of equations (4.50) is solvable for $\alpha_1, \alpha_2, \ldots, \alpha_m$ non trivially or not.

To answer this question, we simply need to subdivide this very discrete case into three subcases:

Subcase: $N = m + 1$:

In this case, the system of equations (4.50) can be rewritten in the following matrix form as follows:

$$\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ x_2 & x_2^2 & x_2^3 & \dots & x_2^m \\ x_3 & x_3^2 & x_3^3 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_m & x_m^2 & x_m^3 & \dots & x_m^m \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4.51}$$

It is clearly seen that the left hand side of the above matrix relation contains a non singular Vandermonde matrix of order $m \times m$. This proves that there exists solely the trivial solution of (4.50), namely

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

which shows that, according to the result (4.45), since no non trivial solution of (4.50) exists, $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is positive definite.

Subcase: $N > m + 1$:

In this case, the system of equations (4.50) has more number of equations than the number of unknowns. Moreover, $x_2, x_3, \dots, x_{N-1}$ are all non zero real numbers.

This can only lead to the conclusion that the solution of (4.50) is just one

and the solution is trivial, namely

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

which again shows that, according to the result (4.45), since no non trivial solution of (4.50) exists, $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is positive definite.

Subcase: $N < m + 1$:
In this case, the system of equations (4.50) has more number of unknowns than the number of equations. Moreover, $x_2, x_3, \ldots, x_{N-1}$ are all non zero real numbers.

This can only lead to the conclusion that the number of non trivial solutions of (4.50) are infinitely many.

We shall show that $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is not positive definite in this case by a simple counter example:

Let us take $N = 3, m = 3$ for $\mathbf{X}_{(3)} = (X, X^2, X^3)$ and $\mathcal{X}_X(\{d\}) = \{0, \frac{1}{2}, 1\}$.

Therefore, one of the solutions of

$$\begin{cases} \alpha_1 + \alpha_2 + \alpha_3 & = 0 \\ \alpha_1 \left(\frac{1}{2}\right) + \alpha_2 \left(\frac{1}{2}\right)^2 + \alpha_3 \left(\frac{1}{2}\right)^3 & = 0 \end{cases} \tag{4.52}$$

is

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{8} \\ -\frac{3}{8} \\ \frac{1}{4} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

which rather shows that

$$P_{X|\{d\}} \left(\left\{ x | \alpha_1 x + \alpha_2 x^2 + \ldots \alpha_m x^m = 0 \right\} \right) = 1$$

and thus $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is not positive definite in this case.

Continuous Case: X is continuous:

In this case, the range of variability of X is $\mathcal{X}_X(\{d\}) = [0, 1]$, which contains infinitely many non zero real values of x.

Therefore, we can safely conclude that for any given value of $m \in \mathbb{N}$, there can only be trivial solutions of (4.50), simply because the non zero values of x are infinitely many.

This only proves that, in accordance with the result (4.45), since no non trivial solution of (4.50) exists, $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is positive definite in this continuous case.

So, by **summarizing** everything, we conclude:

1. If X is discrete, then $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is **positive definite** for $N \geq m + 1$
2. If X is continuous, then $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is **always positive definite** □

Notably, the **positive definite** covariance matrix $\mathsf{Cov}[\mathbf{X}_{(m)}]$ (for $N \geq m+1$ in the discrete cases) has been symbolized alternatively (abbreviated for the sake of better clarity) by $(\sigma_{i,j})_{1 \leq i,j \leq m}$ described by (4.43).

Whence, with the **formal proof of the positive definiteness** of $\mathbf{f}'(\overrightarrow{\boldsymbol{\beta}}) = \mathsf{Cov}[\mathbf{X}_{(\boldsymbol{m})}]$ for every $\overrightarrow{\boldsymbol{\beta}} \in \mathbb{R}^m$ the **proposition 4.5.2** is thereby proved. □

4.5.4 The formulation of the solvability criterion

Proposition 4.5.3 (The solvability criterion). *The criterion for solvability is plainly the* ***invertibility*** *of the matrix* $\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}})$ *for every* $\overrightarrow{\boldsymbol{\beta}} \in \mathbb{R}^m$. *In other words, the matrix* $\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}})\right)^{-1}$ ***must*** *exist for every* $\overrightarrow{\boldsymbol{\beta}} \in \mathbb{R}^m$.

Equivalently*, this solvability criterion is basically nothing, but the* ***non-singularity*** *of the matrix* $\mathsf{Cov}[\mathbf{X}_{(m)}]$.

Proof of the ***proposition 4.5.3****.* Notably, any covariance matrix is **positive semi-definite** anyway and therefore the **non-singularity** of the covariance matrix $\mathsf{Cov}[\mathbf{X}_{(m)}]$ necessitates the **positive definiteness** of the same.

The proof of this proposition took exactly two steps and in fact by proving the **proposition 4.5.1** (given in the subsection 4.5.2), namely $\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}}) = \mathsf{Cov}[\mathbf{X}_{(\boldsymbol{m})}]$ for every $\overrightarrow{\boldsymbol{\beta}} \in \mathbb{R}^m$ **at first** and then proving the **proposition 4.5.2** (given in the subsection 4.5.3), namely the **positive definiteness** of $\mathsf{Cov}[\mathbf{X}_{(m)}]$ **thereafter**. The proof of this very positive definiteness therefore proves the **proposition 4.5.3**. □

Corollary 4.5.1 (Usability of the Newton Raphson procedure). *Because of the positive definiteness of* $\mathsf{Cov}[\mathbf{X}_{(m)}]$ *for every* $\overrightarrow{\boldsymbol{\beta}} \in \mathbb{R}^m$, $\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}})$ *is* ***invertible anyway*** *and thus, the* $\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\boldsymbol{\beta}})\right)^{-1}$ *is* ***determinable anyway*** *and thereby establishing the* ***solvability criterion*** *for the usage of the Newton Raphson procedure.*

Remark 4.5.1. *Notably, the* ***Banach's fixed point theorem*** *(the elaboration and the statement of the theorem are referred to the pages from 148 to 150 of [5]) for the determination of a zero of a multivariate function (denoted by* $\overrightarrow{\mathbf{f}}$ *in our case), which describes the* ***convergence radius*** *of the solution* $\overrightarrow{\boldsymbol{\beta}}^*$, *necessitates* ***this very solvability criterion*** *as well.*

4.5.5 Special consideration of the cases for $m \in \{0, 1, 2\}$

These special cases, viz. $m \in \{0, 1, 2\}$, as we know, are of special importance, as far as the consideration of the standard minimum information probability distributions are concerned. However, the case for $m = 0$ is too trivial and therefore does not require any needful discussion.

The solvability criteria, namely the positive definiteness of both the matrices $\mathsf{Cov}[\mathbf{X}_{(1)}]$ and $\mathsf{Cov}[\mathbf{X}_{(2)}]$ play a principle role in programming the computation procedures of minimum information monotone and uni- extremal probability distributions.

In this subsection, we shall basically consider the very fact that the problems arising out of **programming an algorithm** differs from a **theoretical problem** to a certain degree and how the possible programming difficulties can be effectively resolved by using the aforesaid solvability criterion.

So, let us consider both the monotone and the uni- extremal cases one by one as follows:

Firstly, in **minimum information monotone cases**,

$$\Delta_\beta = \det\left(\mathsf{Cov}[\mathbf{X}_{(1)}]\right) = \frac{d\mu_1}{d\beta} = \mu_2 - \mu_1^2 > 0$$

with reference to the subsection 6.2.1, we have set $\mu_i = \mu_i^{(\beta)}$ as a function of β $(\neq 0)$ for $i \in \{1, 2\}$, as described in (12.105).

In each Newtonian step of the Newton Raphson procedure, β is incremented by a quantity $h = -\frac{\mu_1 - \mu_1^*}{\mu_2 - \mu_1^2}$.

Purely from the theoretical point of view, which shall be established in due course, $\mu_2 - \mu_1^2 = 0 \Leftrightarrow |\beta| = +\infty$.

Though the solvability criteria **theoretically** says that $\mu_2 - \mu_1^2 > 0$ for **every finite value** of β, the programming difficulty is not completely unresolved, if a numerical computation of $\mu_2 - \mu_1^2$ coincidentally turns out to be 0, which would **not enable** the computation of h finitely. This could happen, only if $|\beta|$ unluckily turns out to be abnormally large in any computational process. As a programmer, I do advise my users to keep the input value of μ_1^* restricted to $0.00001 \leq \mu_1^* \leq 0.9999$, though $0 < \mu_1^* < 1$ is **particularly** the theoretical restriction for μ_1^*.

In **discrete cases**, I would advise my users to restrict N preferably to $N \leq 10000$ in general. Should the user intends to input $\mu_1^* = 0.9999$, then I would advise the restriction of N to $N \leq 2500$.

The trivial case of $\beta = 0$ implies and implied by $\mu_1 = \frac{1}{2}$ and therefore no Newton Raphson procedure is of necessity then.

Secondly, in **minimum information uni-extremal cases**,

$$\Delta_{\beta\gamma} = \det\left(\mathsf{Cov}[\mathbf{X}_{(2)}]\right) = \begin{vmatrix} \frac{\partial \mu_1}{\partial \beta} & \frac{\partial \mu_1}{\partial \gamma} \\ \frac{\partial \mu_2}{\partial \beta} & \frac{\partial \mu_2}{\partial \gamma} \end{vmatrix} = (\mu_4 - \mu_2^2)(\mu_2 - \mu_1^2) - (\mu_3 - \mu_1\mu_2)^2 > 0$$

with reference to the subsection 6.3.9, we have set $\mu_i = \mu_i^{(\beta,\gamma)}$ as a function simultaneously of β and γ ($\neq 0$) for $i \in \{1, 2, 3, 4\}$, as described in (12.107).

For programming (practical) reasons, the system of equations (4.4) for $m = 2$ is **conveniently equivalently** rewritten, as described in (12.95) and (12.96), as

$$\begin{cases} \int\limits_{\mathscr{X}_X} (x - \mu_1^*) e^{\beta x + \gamma x^2} \nu_X(dx) = 0 \\ \int\limits_{\mathscr{X}_X} (x^2 - \mu_2^*) e^{\beta x + \gamma x^2} \nu_X(dx) = 0 \end{cases}$$

for the purpose of solving for (β, γ).

In each Newtonian step, β and γ are incremented by h and k respectively (the description is referred to (12.111)).

The **theoretical positivity** of $\Delta_{\beta\gamma}$ is hereby utilized to make sure that, the real value of the denominator Ψ given in (12.114) is **ought** to be **non-zero**, thereby enabling h and k to be **finitely computable**.

The derivation of Ψ given in (12.114) in terms of $\Delta_{\beta\gamma}$ can be easily shown as

$$\Psi = \begin{vmatrix} \mu_2 - \mu_1^*\mu_1 & \mu_3 - \mu_1^*\mu_2 \\ \mu_3 - \mu_1\mu_2^* & \mu_4 - \mu_2^*\mu_2 \end{vmatrix}$$

$$= \Delta_{\beta\gamma} + \mu_1 \begin{vmatrix} \mu_1 - \mu_1^* & \mu_3 - \mu_1\mu_2 \\ \mu_2 - \mu_2^* & \mu_4 - \mu_2^2 \end{vmatrix} + \mu_2 \begin{vmatrix} \mu_2 - \mu_2^* & \mu_3 - \mu_1\mu_2 \\ \mu_1 - \mu_1^* & \mu_2 - \mu_1^2 \end{vmatrix}$$

Therefore, the **legitimate smallness** of both the expressions $|\mu_1 - \mu_1^*|$ and $|\mu_2 - \mu_2^*|$ **ensures the strict positivity** of Ψ (because of the strict positivity of $\Delta_{\beta\gamma}$) and thereby resolving the possible (practical) programming problem.

Though the debugging of both the software programs meant for computing the uni- extremal probability distributions (i.e. in both discrete and continuous cases) has **never** shown that Ψ could **ever** be **computationally** zero (viz. $\Psi = 0$), the issue with regard to the programming quality must be kept in mind. As a programmer, I have imperatively taken care of this aforesaid smallness skillfully.

Even if a particular course of computation of h or k encounters $\Psi = 0$, **still** the Newton Raphson procedure would not cease to continue, but may delay delivering the final result to a marginally negligibly small extent only.

The special case of $\gamma = 0$ regards to the immediately discussed preceding minimum information monotone case.

4.6 The special Hankel matrix

Another name of a Hankel matrix is Hankel kernel. In this section, we principally intend to show that the **positive definiteness** of the Hankel matrix (4.23), namely the matrix

$$\begin{pmatrix} 1 & \mu_1 & \mu_2 & \dots & \mu_n \\ \mu_1 & \mu_2 & \mu_3 & \dots & \mu_{n+1} \\ \mu_2 & \mu_3 & \mu_4 & \dots & \mu_{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_n & \mu_{n+1} & \mu_{n+2} & \dots & \mu_{2n} \end{pmatrix} \tag{4.23}$$

plays a predominant role in the following:

1. The **aforesaid solvability criterion**. The positive definiteness of the Hankel matrix (4.23) **implies and implied by** the positive definiteness of the covariance matrix $\mathsf{Cov}[\mathbf{X}_{(n)}] = (\sigma_{i,j})_{1\leq i\,,\,j\leq n}$ given by (4.43) for $n \in \mathbb{N}$.

2. Proving the **uniqueness** of the solution of the system of equations (4.4) (or equivalently of 4.2). The uniqueness of the solution of (4.4) has been **reproved** in the **appendix** B by using the positive definiteness of the Hankel matrix (4.23).

Therefore, let us prove the positive definiteness of the Hankel matrix (4.23) by using the positive definiteness of the covariance matrix $\mathsf{Cov}[\mathbf{X}_{(n)}]$ for $n \in \mathbb{N}$.

Proposition 4.6.1. *The Hankel matrix (4.23) is* ***positive definite****.*

Proof of the ***proposition 4.6.1****.* The determinant of the Hankel matrix (4.23) is equal to

$$\begin{vmatrix} 1 & \mu_1 & \mu_2 & \dots & \mu_n \\ \mu_1 & \mu_2 & \mu_3 & \dots & \mu_{n+1} \\ \mu_2 & \mu_3 & \mu_4 & \dots & \mu_{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_n & \mu_{n+1} & \mu_{n+2} & \dots & \mu_{2n} \end{vmatrix}$$

$$= \begin{vmatrix} 1 & \mu_1 & \mu_2 & \dots & \mu_n \\ \mu_1 - \mu_1(1) & \mu_2 - \mu_1(\mu_1) & \mu_3 - \mu_1(\mu_2) & \dots & \mu_{n+1} - \mu_1(\mu_n) \\ \mu_2 - \mu_2(1) & \mu_3 - \mu_2(\mu_1) & \mu_4 - \mu_2(\mu_2) & \dots & \mu_{n+2} - \mu_2(\mu_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_n - \mu_n(1) & \mu_{n+1} - \mu_n(\mu_1) & \mu_{n+2} - \mu_n(\mu_2) & \dots & \mu_{2n} - \mu_n(\mu_n) \end{vmatrix}$$

where the i^{th} **modified** row $R_i^{'}$
$= i^{th}$**original** row $R_i - (\mu_{i-1} \times$ **first** row $R_1)$ for $i \in \{2, 3, \dots, n+1\}$

$$= \begin{vmatrix} 1 & \mu_1 & \mu_2 & \dots & \mu_n \\ 0 & \mu_2 - \mu_1(\mu_1) & \mu_3 - \mu_1(\mu_2) & \dots & \mu_{n+1} - \mu_1(\mu_n) \\ 0 & \mu_3 - \mu_2(\mu_1) & \mu_4 - \mu_2(\mu_2) & \dots & \mu_{n+2} - \mu_2(\mu_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \mu_{n+1} - \mu_n(\mu_1) & \mu_{n+2} - \mu_n(\mu_2) & \dots & \mu_{2n} - \mu_n(\mu_n) \end{vmatrix}$$

$$= \begin{vmatrix} \mu_2 - \mu_1(\mu_1) & \mu_3 - \mu_1(\mu_2) & \dots & \mu_{n+1} - \mu_1(\mu_n) \\ \mu_3 - \mu_2(\mu_1) & \mu_4 - \mu_2(\mu_2) & \dots & \mu_{n+2} - \mu_2(\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n+1} - \mu_n(\mu_1) & \mu_{n+2} - \mu_n(\mu_2) & \dots & \mu_{2n} - \mu_n(\mu_n) \end{vmatrix}$$

$$= \begin{vmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_1}{\partial \beta_2} & \dots & \frac{\partial \mu_1}{\partial \beta_n} \\ \frac{\partial \mu_2}{\partial \beta_1} & \frac{\partial \mu_2}{\partial \beta_2} & \dots & \frac{\partial \mu_2}{\partial \beta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_n}{\partial \beta_1} & \frac{\partial \mu_n}{\partial \beta_2} & \dots & \frac{\partial \mu_n}{\partial \beta_n} \end{vmatrix} \quad (\text{ by (4.42) })$$

$$= \det\left(\mathsf{Cov}[\mathbf{X}_{(n)}]\right) > 0$$

□

which proves the **positive definiteness** of the Hankel matrix (4.23) (i.e. the proposition 4.6.1) and thereby establishing our aforesaid **two assertions**.

□

Chapter 5

Suitable simple bounds of moments

Since this dissertation is confined to the probability distributions of **compact supports** only, any moment of such a probability distribution **must** have a **bounded range of variability**. In this chapter, we shall discuss about the importance of such **bounds**.

For the **sake of simplicity**, as usual, **without any loss of generality**, we shall basically confine our discussions to the bounds of **every** moment μ_n of the random variable X (i.e. for **every** $n \in \mathbb{N}$), by keeping the **statement 2.4.2** in mind (the statement describing the **indexing of moments**). We shall discuss about the bounds of the **standard** moments of the random variable Y thereafter, which are **relevant** for this dissertation.

5.1 Basic statements for the moments of X

Statement 5.1.1 (The statement on insignificance). *Referring to our joint paper [39], the higher moments of X* ***do not really deliver*** *any additional needful* ***quantitative information*** *for the purpose of constructing the targeted approximating probability distribution (namely $f_{X|\{d\}}(x)$, $x \in \mathcal{X}_X$) of X. In this very sense, these higher moments of X are* ***insignificant****. This* ***insignificance*** *of the moment μ_n, $n \in \mathbb{N}$ is ascribed to the* ***smallness*** *of the* ***range of variability*** *(referred to the* ***definition 5.2.5****) of μ_n. However, for practical reasons, the* ***smallness*** *of the* ***bounded range*** *(referred to the* ***definition 5.2.6****) of μ_n is under our consideration.*

Statement 5.1.2 (The restrictions on moments of X). *If X happens to be **continuous**, then the **least upper bound** (denoted by $lub(\mu_i)$) and the **greatest lower bound** (denoted by $glb(\mu_i)$) of the i^{th} moment μ_i (of X) for each and every $i \in \{1, 2, \ldots, m\}$ can be determined by the usage of the **positive definiteness of Hankel matrices** defined in the subsection (4.4.2) and in fact,*

- *the positive definiteness of the Hankel matrices given by (4.23) and (4.30) determines the **greatest lower** and **least upper** bounds of μ_i for an **even** i $(= 2n)$, $n \in \{1, 2, \ldots, \lfloor \frac{m}{2} \rfloor\}$*
- *the positive definiteness of the Hankel matrices given by (4.27) and (4.33) determines the **greatest lower** and **least upper** bounds of μ_i for an **odd** i $(= 2n + 1)$, $n \in \{0, 1, 2, \ldots, \lfloor \frac{m-1}{2} \rfloor\}$*

*This aforesaid positive definiteness is obviously a **severe restriction** imposed on the variability of μ_i.*

***In fact**, for the **discrete** case of X, the **aforesaid restriction** on moments of X is **even severer**. This severeness of the restriction of the range of variability of μ_i is **well explained** by the very fact that the support $\mathcal{X}_X = \{x_1, x_2, \ldots, x_N\}$ is **predetermined** and therefore the range of variability of μ_i depends **additionally** on the choice of $\mathcal{X}_X$.*

The number of moments of X needed to construct the approximating probability distribution of X is importantly a big question with regard to the practicability. A **subsequent** statement in this very regard is given as follows:

Statement 5.1.3 (Practicability pertaining to the number of moments). *Every additional moment of X needed to construct the exponential polynomial probability distribution of X means that the amount of **numerical mathematical work** as well as the amount of **programming work** increases **exponentially**.*

*Therefore, even if the **exact values** of the moments of X are **hypothetically assumed to be known**, barely from the **practical point of view**, one must make sure that only the **needed** number of moments of X should be included, which are unavoidably necessary.*

We shall discuss both the **discrete** and the **continuous** cases of X, starting with the continuous case though.

5.2 The formulation of the first target

The formulation of the first target pertains to the **continuous** case of the random variable X.

5.2.1 The background

In the subsection 3.4, we have already established that the probability density $f_{X|\{d^{(n)}\}}(x) = e^{\beta_0 + \sum_{i=1}^{n} \beta_i x^i}$, $0 \leq x \leq 1$, $n \in \mathbb{N}$ is a **consistent** density estimator of the unknown **situation oriented need based** probability density $f_X(x)$, $0 \leq x \leq 1$, with

- $\beta_0 = \log\left(\int_0^1 e^{\sum_{i=1}^{n} \beta_i x^i} dx\right)$
- $d^{(n)} = (\mu_1, \mu_2, \ldots, \mu_n)$

In other words, the **goodness of the approximation** of the aforesaid probability density $f_X(x)$ (approximated by the aforesaid probability density $f_{X|\{d^{(n)}\}}(x)$) can be **improved arbitrarily**, by increasing the number of moments n.

Therefore, the question arises, how exactly the moments do **behave** for to contribute to the desired goodness of this approximation? This very question is the question of insignificance of certain moments (referred to the **statement 5.1.1**). It is well **intuitively assertible** the **range of variability** of μ_n defined by the open interval $(glb(\mu_n), lub(\mu_n))$ becomes **smaller and smaller** with the increase in n.

So, the moment μ_n with a smaller range of validity has understandably a **lesser strength of being informative**, if the length $lub(\mu_n) - glb(\mu_n)$ of the open interval $(glb(\mu_n), lub(\mu_n))$ happens to be **smaller**.

This **smallness** of $lub(\mu_n) - glb(\mu_n)$ principally says that the inclusion of the moment μ_n for the purpose of constructing the approximating probability density $f_{X|\{d\}}(x)$ **improves** the approximation of the probability density $f_X(x)$ only to a **negligible** extent.

In other words, if this interval length $lub(\mu_n) - glb(\mu_n)$ is **insignificant enough**, then it **hardly matters**, whether the approximating probability

density could be accepted as $f_{X|\{d^{(n-1)}\}}(x)$ (by taking $d^{(n-1)} = (\mu_1, \mu_2, \ldots, \mu_{n-1})$ by **excluding** the moment μ_n) or could be accepted as $f_{X|\{d^{(n)}\}}(x)$ (by taking $d^{(n)} = (\mu_1, \mu_2, \ldots, \mu_n)$ by **including** the moment μ_n).

So, the smallness of $lub(\mu_n) - glb(\mu_n)$ particularly becomes a **decision making factor**, whether the **inclusion** of μ_n **really matters or not**.

5.2.2 The first target

The **core** of the discussion is about the very fact that only a **finite** number of moments of X are **enough** to construct the targeted **probability density function** $f_{X|\{d\}}(x), x \in [0, 1]$ of X **appropriately**. Now, by targeting the smallness of $lub(\mu_n) - glb(\mu_n)$, we must find out a **method** of choosing an **optimal value**[1] of n, so that a predeterminately chosen **legitimate smallness**[2] of $lub(\mu_n) - glb(\mu_n)$ is duly achieved? In other words, if the predeterminate choice of the said smallness is denoted by a small positive number ϵ, what should be the **minimum value** of n, so as to fulfill the condition $lub(\mu_n) - glb(\mu_n) \leq \epsilon$?

Unfortunately, the exact determination of $glb(\mu_n)$ (by means of the positive definiteness of the Hankel matrices defined by **(4.23)** or **(4.27)**) and of $lub(\mu_n)$ (by means of the positive definiteness of the Hankel matrices defined by **(4.30)** or **(4.33)**) for every $n \in \mathbb{N}$ happens to be **extremely cumbersome**.

So, we shall have to look for an alternative method giving an **optimally chosen value** of n. This **optimal value** of n, with regard to the **statement 5.1.3**, should be as **small as possible**.

Accordingly, we shall choose a **legitimate lower bound** of μ_n (denoted by $lb(\mu_n)$), such that $lb(\mu_n) \leq glb(\mu_n)$ for every $n \in \mathbb{N}$) as well as a **legitimate upper bound** of μ_n (denoted by $ub(\mu_n)$), such that $ub(\mu_n) \geq lub(\mu_n)$ for every $n \in \mathbb{N}$) with subject to the following two **legitimate prerequisites**:

1. Since **each** of $glb(\mu_n)$ and $lub(\mu_n)$ depends on the lower ordered moments namely $\mu_1, \mu_2, \ldots, \mu_{n-1}$, then each of $lb(\mu_n)$ and $ub(\mu_n)$ **must** depend on $\mu_1, \mu_2, \ldots, \mu_{n-1}$ as well.

[1] Optimality in this regard is referred to **minimality**.

[2] the legitimacy of this smallness is regarded to the problem-related issue.

In other words, both $lb(\mu_i)$ and $ub(\mu_i)$ **must** depend **at least** on μ_{i-1}, for every $i \in \{2, 3, \ldots\}$.

2. For the standard cases, namely for the **minimum information monotone cases** (i.e. cases for $n = 1$) and for the **minimum information uni-extremal cases** (i.e. cases for $n = 2$), we **must** have $lb(\mu_n) = glb(\mu_n)$ and $ub(\mu_n) = lub(\mu_n)$.

Obviously, the **minimum** value of n, which fulfills $lub(\mu_n) - glb(\mu_n) \leq \epsilon$ is **smaller** than that which fulfills $ub(\mu_n) - lb(\mu_n) \leq \epsilon$.

But, as already discussed, because of the **complexities** of the expressions of both $glb(\mu_n)$ and $lub(\mu_n)$, we **do not have any better choice** at the moment other than determining suitable expressions of $lb(\mu_n)$ and $ub(\mu_n)$ that are **simpler** in nature.

Accordingly, on defining the suitably chosen expressions of $lb(\mu_n)$ and $ub(\mu_n)$, we shall denote the **optimally smallest possible** value (i.e. the **optimal** value) of n fulfilling the condition $ub(\mu_n) - lb(\mu_n) \leq \epsilon$ as $\mathbf{N_{optimal}}$.

In order to make sure that $\mathbf{N_{optimal}}$ is **actually** the **smallest** value of n described by

$$ub(\mu_{\mathbf{N_{optimal}}}) - lb(\mu_{\mathbf{N_{optimal}}}) \leq \epsilon \tag{5.1}$$

$$ub(\mu_{\mathbf{N_{optimal}}-1}) - lb(\mu_{\mathbf{N_{optimal}}-1}) > \epsilon \tag{5.2}$$

we must prove that the sequence $\{ub(\mu_n) - lb(\mu_n)\}_{n\in\mathbb{N}}$ must be **strictly monotonically decreasing** and **converges** to 0 (the proof is referred to the proposition 5.4.1). This is the **first target** of this chapter, after having defined $lb(\mu_n)$ and $ub(\mu_n)$ suitably, **for the purpose** picturing the value of $\mathbf{N_{optimal}}$ (the value of $\mathbf{N_{optimal}}$ regards **mainly** to the **continuous cases** of X **only**).

So, by keeping the above stated arguments in mind, we introduce the following definitions:

Definition 5.2.1 (The $lb(\mu_n)$, $n \in \{1, 2, \ldots, m\}$). $lb(\mu_n) = \frac{\mu_{n-1}^2}{\mu_{n-2}}$
Notably, it shall be established in due course that, theoretically $\mu_{-1} = \infty$ for both discrete and continuous cases, so as to establish $lb(\mu_1) = 0$.

Definition 5.2.2 (The $ub(\mu_n)$, $n \in \{1, 2, \ldots, m\}$). $ub(\mu_n) = \mu_{n-1}$.

Definition 5.2.3 (Optimal value of n). *As we have already discussed, the optimal value of* $\boldsymbol{n}$ *is defined to be the* ***minimum*** *value of* $\boldsymbol{n}$*, which fulfills the condition* $ub(\mu_n) - lb(\mu_n) \leq \epsilon$*. As in (5.1), the optimal value of* $\boldsymbol{n}$ *is symbolized as* $\mathbf{N_{optimal}}$

Definition 5.2.4 (Optimal choice of m). *As we know, the index* $\boldsymbol{m}$ *stands for the* ***exact*** *number of moments needed to construct the approximating probability distribution of* X*.*

The consideration of (5.1) says that, the ***inclusion*** *of the moment* $\mu_{\mathbf{N_{optimal}}}$ *(because of the* ***inadmissible smallness*** *of* $ub(\mu_{\mathbf{N_{optimal}}}) - lb(\mu_{\mathbf{N_{optimal}}})$ *) is simply* ***unjustified****. Therefore* $\mu_{\mathbf{N_{optimal}}}$ *becomes* ***redundant****.*

But*, the consideration of (5.2) says that,* $ub(\mu_{\mathbf{N_{optimal}}-1}) - lb(\mu_{\mathbf{N_{optimal}}-1})$ *is* ***just not small enough*** *for the moment* $\mu_{\mathbf{N_{optimal}}-1}$ *to be redundant.*

Hence, the optimal choice of $\boldsymbol{m}$ *is* $\mathbf{N_{optimal}} - 1$*, i.e.* $\boldsymbol{m} = \mathbf{N_{optimal}} - 1$*.*

Definition 5.2.5 (Range of variability). *The range of variability of the moment* μ_n *is the open interval* $(glb(\mu_n), lub(\mu_n))$*. The* ***necessary and the sufficient*** *condition for the existence of the probability distribution of* X *has been stated by the* ***theorem 4.4.1*** *given in the* ***subsection 4.4.2*** *says that* $glb(\mu_n) < \mu_n < lub(\mu_n)$ *for every* $n \in \{1, 2, \ldots, m\}$ *implies and implied by the* ***existence of the probability distribution of*** X*.*

Definition 5.2.6 (Bounded range). *The bounded range of the moment* μ_n *is the open interval* $(lb(\mu_n), ub(\mu_n))$*. The* ***necessary*** *condition for the existence of the probability distribution of* X *is given by* $lb(\mu_n) < \mu_n < ub(\mu_n)$ *for every* $n \in \{1, 2, \ldots, m\}$*. Thus, if* μ_n ***exceeds*** *its upper bound* $ub(\mu_n)$ *or* ***falls below*** *its lower bound* $lb(\mu_n)$*, then the* ***existence*** *of the probability distribution of* X *(i.e. the* ***existence*** *of the* ***solution of the system of equations (4.4)****) is* ***completely ruled out****.*

Remark 5.2.1. *The* ***bounded range*** *of* μ_n *is undoubtedly the* ***superset*** *of the* ***range of validity*** *of* μ_n*.*

Remark 5.2.2. *Throughout this dissertation, we shall assume that any choice of* $\boldsymbol{m}$ *shall be dependent on the choice of* $\boldsymbol{n}$*.*

5.3 The formulation of the second target

The formulation of the second target pertains to the **discrete** case of the random variable X.

5.3.1 The background

The expressions of $lb(\mu_n)$ and $ub(\mu_n)$ proposed by the definitions 5.2.1 and 5.2.2 **are taken for discussions** in the discrete cases of X **as well** (i.e. as that of the continuous cases of X). We shall show that

- $glb(\mu_1) = lb(\mu_1)$ holds, but $glb(\mu_2) = lb(\mu_2)$ **does not necessarily hold** and in fact $glb(\mu_2) \geq lb(\mu_2)$.
- $lub(\mu_n) = ub(\mu_n)$ holds for $n \in \{1, 2\}$.

Unlike the continuous cases of X, the **theorem 4.4.1 does not necessarily apply** for the discrete cases of X, simply because in certain cases $glb(\mu_2) > lb(\mu_2) = \mu_1^2$ **must** hold. That is, the $glb(\mu_n)$, **in general, is not determinable** by the positive definiteness of the Hankel matrices defined in (4.23) (or (4.27)). However, the $lub(\mu_n)$ **seems to be determinable** by the positive definiteness of the Hankel matrices defined in (4.30) (or (4.33)). This very issue, which refers to the **statement 5.1.2**, shall be discussed in this chapter.

Unlike the continuous cases of X, with reference to the **exact representation** of the probability mass function of X (in the **discrete cases** of X) stated by (2.9), the **optimal** value $\mathbf{N_{optimal}} - 1$ of m can **never** be greater than $N - 1$ (N being the number of (discrete) elements of the support of the probability distribution of X).

In other words, $m = \mathbf{N_{optimal}} - 1 \leq N - 1$, i.e. $\mathbf{N_{optimal}} \leq N$.

Thus, with subject to **any** predeterminate choice of ϵ (**as in the continuous case** of X), the condition $ub(\mu_n) - lb(\mu_n) \leq \epsilon$ **may** or **may not** be fulfilled, if the value $\mathbf{N_{optimal}}$ is **constrained to** $\mathbf{N_{optimal}} \leq N$.

5.3.2 The second target

Therefore, for a **discrete** X, we shall have to take care of the following:

- the detailed discussions about $glb(\mu_1) = lb(\mu_1) = 0$ and $lub(\mu_1) = ub(\mu_1) = 1$ are **relatively easy** and are carried out in the **subsection 6.2.3** concerning the existence of the discrete monotonic probability distribution of X.

- the detailed discussions about $glb(\mu_2) \geq lb(\mu_2) = \mu_1^2$ and $lub(\mu_2) = ub(\mu_2) = \mu_1$ need a very **careful attention**. So, this leads us to think that $glb(\mu_n)$ for every $n \in \mathbf{N}$ is not determinable by the positive definiteness of the Hankel matrices 4.23 (or 4.27). Establishing this picture is the **second target** of this chapter.

- the discussions about $lb(\mu_n)$ and $ub(\mu_n)$ for $n \geq 3$ with the aim of evaluating the $\mathbf{N_{optimal}}$ may be **largely analogous** to the same for the **continuous cases** of X. These detailed discussions, i.e. for the higher moments, namely $\mu_3, \mu_4, \ldots$ etc. are avoided in this chapter.

  As a matter of fact, the higher moments $\mu_3, \mu_4, \ldots$ have, as we have already mentioned, **no practical relevance** in the field of stochastic science. Obviously, these discussions are rather complicated.

As the next step, we shall establish the significance of the **definitions 5.2.1 and 5.2.2** in form of proving the inequality $lb(\mu_n) < \mu_n < ub(\mu_n)$ for every $n \in \mathbb{N}$.

5.4 Essential role of the bounded range

In this section, we shall discuss the most deciding role of the bounding range defined by the interval $(lb(\mu_n)\,,\, ub(\mu_n)$.

In case X happens to be discrete, the trivial cases for $N \leq 2$ could be trivially kept out of consideration as these trivial cases basically refer to **degenerated** or **bernoulli** probability distributions. The consideration of the same shall however be carried out in due course.

The proof of the **theorem 5.4.2** meant for establishing the inequalities $lb(\mu_n) < \mu_n$ and $\mu_n < ub(\mu_n)$ (in accordance with the **definition 5.2.6** of the bounded range), the usage of the well known and well used **Cauchy-Schwarz inequality** in the **probability theory** is necessary and is therefore introduced as follows:

Theorem 5.4.1 (Cauchy-Schwarz inequality). *If X_1 and X_2 be the two random variables, each having the same range of variability, then*

$$|E[X_1X_2]| \leq \sqrt{E[X_1^2]}\sqrt{E[X_2^2]} \tag{5.3}$$

whereas $|E[X_1X_2]| = \sqrt{E[X_1^2]}\sqrt{E[X_2^2]}$ holds, if X_1 and X_2 are correlated in a way that $\exists\ \kappa \in \mathbb{R}$, such that $X_1 = \kappa X_2$.

The proof of the **theorem 5.4.1** is referred to [8].

5.4.1 The inequality defining the bounded range

Theorem 5.4.2 (The bounding inequality). *For every moment of X denoted by μ_n, $n \in \mathbb{N}$ the following inequality holds:*

$$lb(\mu_n) = \frac{\mu_{n-1}^2}{\mu_{n-2}} < \mu_n < \mu_{n-1} = ub(\mu_n), \qquad n \in \mathbb{N} \tag{5.4}$$

In case X happens to be discrete, then the above inequality (5.4) is principally taken for $N \geq 3$.

Proof of the theorem 5.4.2. For the general proof of the theorem 5.4.2, namely the proof of the inequality (5.4), let us examine the particular cases for $n \in \{1, 2\}$ at first.

Only after this, we shall proceed to prove the theorem for any $n \in N$, i.e. including the cases $n \geq 3$.

By taking

$$p_X(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_m x^m \tag{5.5}$$

we can prove the above inequality for $n \in \{1, 2\}$ trivially as follows:

The theoretical calculation of μ_{-1} shows that

- for X being discrete, $\mu_{-1} = \sum_{j=1}^{N} x_j^{-1} e^{p_X(x_j)} = \infty$, as $x_1 = 0$
- for X being continuous, $\mu_{-1} = \int_0^1 x^{-1} e^{p_X(x)} dx = \infty$, as for the aforesaid polynomial $p_X(x)$, the integral $\int_0^1 x^{-1} e^{p_X(x)} dx$ **diverges**. This divergence can be easily shown by a test named μ-**test for converge or divergence of improper integrals** (referred to the page 254 of [17])

This brings us to the very fact that, for $\mu_0 = 1$,

$$0 = \frac{\mu_0^2}{\mu_{-1}} < \mu_1 < \mu_0 = 1$$

which was our basic assumption right from the beginning. This proves the inequality (5.4) for $n = 1$.

Notably, as we have already discussed, $glb(\mu_1) = lb(\mu_1) = 0$ and $lub(\mu_1) = ub(\mu_1) = 1$ do hold for **both continuous and discrete cases** of X.

Again, for $n = 2$, we have nothing, but our already proven following result

$$\mu_1^2 = \frac{\mu_1^2}{\mu_0} < \mu_2 < \mu_1$$

which therefore proves the inequality (5.4) for $n = 2$.

Notably, $glb(\mu_2) = lb(\mu_2) = \mu_1^2$ and $lub(\mu_2) = ub(\mu_2) = \mu_1$ do hold for **continuous** cases of X. But, as we have already mentioned, we shall show in this very chapter that $glb(\mu_2) \geq lb(\mu_2) = \mu_1^2$ and $lub(\mu_2) = ub(\mu_2) = \mu_1$ do hold for **discrete** cases of X.

Now, for the sake of proving the aforesaid inequality (5.4) that includes the cases for $n \geq 3$ with the help the theorem 5.4.1, we proceed as follows:

By taking the range of variability of X_1 and X_2 to be $\mathcal{X}_X$, such that $X_1 = X^{\frac{n}{2}}$ and $X_2 = X^{\frac{n}{2}-1}$, by the inequality (5.3), we easily get

$$\begin{aligned} \left|E\left[X^{n-1}\right]\right| &\leq \sqrt{E\left[X^n\right]}\sqrt{E\left[X^{n-2}\right]} \\ \implies |\mu_{n-1}| &\leq \sqrt{\mu_n}\sqrt{\mu_{n-2}} \end{aligned} \tag{5.6}$$

Here, with subject to

- $N \geq 3$, **if** X is discrete
- or **if** X is continuous

there **cannot** exist such κ for the validity of the **equality relationship** present in the inequality (5.6) and therefore the said inequality (i.e. (5.6)) can be generally (i.e. also for $n \leq 2$) is to be conveniently remodified as

$$\frac{\mu_{n-1}^2}{\mu_{n-2}} < \mu_n \text{ for every } n \in \{1, 2, \ldots, m\} \tag{5.7}$$

(i.e. excluding the possibility $\frac{\mu_{n-1}^2}{\mu_{n-2}} = \mu_n$)

Again, with subject to

- $N \geq 3$, **if** X is discrete
- or **if** X is continuous

we know well that the following holds good:

$$1 > \mu_1 > \mu_2 > \ldots > \mu_n > \ldots \tag{5.8}$$

and this very thing establishes the following:

$$\mu_n < \mu_{n-1} \text{ for every } n \in \mathbb{N} \tag{5.9}$$

Whence, by combining (5.7) and (5.9), we the desired inequality (5.4) gets proven and thereby the **theorem 5.4.2** is proved. □

5.4.2 The monotonic character of the bounded range

The **bounded range** of the moment μ_n, $n \in \mathbb{N}_0$ defined by the **definition 5.2.6** is nothing but the **open interval** $(lb(\mu_n), ub(\mu_n)) = \left(\frac{\mu_{n-1}^2}{\mu_{n-2}}, \mu_{n-1}\right)$.

In this subsection we shall prove that the length of the interval of the bounded range of the moment μ_n, $n \in \mathbb{N}$ **decreases strictly monotonically** with the monotonic increase in n. In order to prove this, we arrive at the following proposition:

Proposition 5.4.1. *The sequence* $\{ub(\mu_n) - lb(\mu_n)\}_{n \in N}$ *is* ***strictly monotonic decreasing*** *and* ***converges to 0****.*

Proof of the ***proposition 5.4.1****.* By rewriting the inequality (5.7), we have

$$\frac{\mu_{n-1}^2}{\mu_{n-2}} < \mu_n \Leftrightarrow \frac{\mu_{n-1}}{\mu_{n-2}} < \frac{\mu_n}{\mu_{n-1}} \Leftrightarrow 1 - \frac{\mu_{n-1}}{\mu_{n-2}} > 1 - \frac{\mu_n}{\mu_{n-1}} \text{ for every } n \in \mathbb{N} \tag{5.10}$$

Thus, by combining the inequalities (5.9) and (5.10), namely $\mu_{n-1} > \mu_n$ and $1 - \frac{\mu_{n-1}}{\mu_{n-2}} > 1 - \frac{\mu_n}{\mu_{n-1}}$ we arrive at

$$\begin{gathered}\underbrace{\mu_{n-1}\left(1 - \frac{\mu_{n-1}}{\mu_{n-2}}\right)}_{=ub(\mu_n)-lb(\mu_n)} > \underbrace{\mu_n\left(1 - \frac{\mu_n}{\mu_{n-1}}\right)}_{=ub(\mu_{n+1})-lb(\mu_{n+1})} \\ \text{i.e.} ub(\mu_n) - lb(\mu_n) > ub(\mu_{n+1}) - lb(\mu_{n+1}) \text{ for every } n \in \mathbb{N}\end{gathered} \tag{5.11}$$

and hence the confirmation of the **strictly monotonic decreasing** character of the sequence $\{ub(\mu_n) - lb(\mu_n)\}_{n\in N}$.

Moreover, since $ub(\mu_n) - lb(\mu_n) < ub(\mu_n) = \mu_{n-1} \to 0$ as $n \to \infty$, that very fact that the sequence $\{ub(\mu_n) - lb(\mu_n)\}_{n\in N}$ **converges to 0**. This proves the **proposition 5.4.1**. □

Corollary 5.4.1 (Optimal value of m). *As an immediate consequence of the proposition 5.4.1 establishing principally the strict monotonic decreasing character of the bounded range of the moment μ_n, $n \in N$, we **reaffirm** and **restate** that $m = \mathbf{N_{optimal}} - 1$ is the **optimal value** of m fulfilling the conditions (5.1) and (5.2).*

Corollary 5.4.2 (Insignificant moments of X). *With subject to a predeterminately given ϵ ($\epsilon > 0$) and in the light of the **fulfillment** of the conditions (5.1) and (5.2) the **insignificant moments** of X are therefore $\mu_{\mathbf{N_{optimal}}}$, $\mu_{\mathbf{N_{optimal}}+1}$, $\mu_{\mathbf{N_{optimal}}+2}$, ... etc. (the concept of **insignificance** is referred to the **statement 5.1.1**).*

Remark 5.4.1 (Infinitesimal character of μ_n for increasing n). *Trivially, the very fact that $\mu_n \to 0$ as $n \to \infty$ can be easily proved by means of the Holder's inequality (see the stated theorem C.1.1 belonging to the appendix section C.1) by taking the functions $f(x) = x^n$ and $g(x) = f_X(x)$ present in the given expression $\mu_n = \int_0^1 x^n f_X(x)dx$ and then by taking the note of $\int_0^1 f(x)dx = \frac{1}{n+1} \to 0$ as $n \to \infty$ thereafter.*

Remark 5.4.2 (Infinitesimal character of the range of variability). *The **range of variability** of the moment μ_n, $n \in \mathbb{N}_0$ defined by the **definition 5.2.5** is nothing but the **open interval** $(glb(\mu_n), lub(\mu_n))$.*

Since, we know that $0 \leq lub(\mu_n) - glb(\mu_n) \leq ub(\mu_n) - lb(\mu_n)$ *and by the proposition 5.4.1,* $\{ub(\mu_n) - lb(\mu_n)\}_{n \in N}$ *is a* ***null sequence*** [3]*, it is obviously clear that* $\{lub(\mu_n) - glb(\mu_n)\}_{n \in N}$ *is too a* ***null sequence****.*

This shows that the range of variability can be made ***infinitesimally*** *small.*

Remark 5.4.3 (The needfulness of the bounded range). *The very fact that the length* $lub(\mu_n) - glb(\mu_n)$ *(of the* ***range of variability*** *of* μ_n*) is strictly monotonically decreasing with the increase in* $\boldsymbol{n}$ *(*$\boldsymbol{n} \in \mathbb{N}$*), is well intuitively assertible, but* ***certainly not*** *easily provable.*

This basically leads us to use the length $ub(\mu_n) - lb(\mu_n)$ *(of the* ***bounded range*** *of* μ_n*), which is strictly monotonically decreasing with the increase in* $\boldsymbol{n}$ *(as shown in the* ***proposition 5.4.1****),* ***instead****.*

Precisely, because of this ***established*** *strictly monotonically decreasing character of* $ub(\mu_n) - lb(\mu_n)$*, we can conclusively state that all the moments* μ_n *for* $\boldsymbol{n} \geq \mathbf{N_{optimal}}$ *(*$\boldsymbol{n} \in \mathbb{N}$*) are* ***insignificant*** *with subject to the predeterminately given* ϵ *(*$\epsilon > 0$*).*

Remark 5.4.4. *However, it remains to be stated that, if we had proven the strictly monotonic decreasing character of* $lub(\mu_n) - glb(\mu_n)$*, then the* ***optimal choice*** *of* $\boldsymbol{m}$ *could have been* ***much lesser*** *than* $\mathbf{N_{optimal}} - 1$*.*

Remark 5.4.5 (Role of a finite number of moments of X). *As a very* ***general statement****, this very strictly monotonically decreasing character of the* ***bounded range*** $ub(\mu_n) - lb(\mu_n)$ *proven by the* ***proposition 5.4.1*** *is* ***not only*** *the case of an* ***exponential polynomial*** *probability density of* X*, but also applicable to the case of* ***any*** *other probability density of* X *that is uniquely determinable by a finite number of moments of* X*. Thus, for a given smallness of* ϵ *(*$\epsilon > 0$*) only a* ***finite number of moments*** *is* ***simply enough*** *to determine a probability density of a random variable.*

5.5 The targeted smallness of the bounded range

Our discussions in this section shall be **principally** confined to the **continuous** cases of X. As we have already mentioned, the analogous arguments

[3] a null sequence is defined to be the sequence that converges to **zero**.

may apply for the **discrete** cases of X largely, but these arguments are **certainly very difficult** (in these discrete cases).

The smallness of the length of the strictly monotonic decreasing bounded range $ub(\mu_n) - lb(\mu_n)$ (of the moment μ_n) (the targeted smallness being described by ϵ) is used to determine the $\mathbf{N_{optimal}}$ (especially in the continuous cases of X).

We have already seen that the $n = \mathbf{N_{optimal}}$ is **uniquely determinable** by appropriate choices of μ_{n-2}, μ_{n-1} as well as of ϵ.

Now, the question arises, with subject to a given chosen value of μ_{n-2}, **what could be** the **worst choice** of μ_{n-1}, as a result of which the condition $ub(\mu_n) - lb(\mu_n) = \mu_{n-1} - \frac{\mu_{n-1}^2}{\mu_{n-2}} \leq \epsilon$ **may** or **may not** be fulfilled? How does this worst choice affects the choice of m?

In order to handle this question in the right way, we arrive at the following proposition:

Proposition 5.5.1 (The worst choice of m). *With subject to a predetermined ϵ and with reference to the statement 5.1.3, **m** should be **as small as possible**. The **worst** possible choice of **m** is dependent exclusively on the **worst** possible choice of μ_{n-1} given by $\mu_{n-1} = \frac{1}{2}\mu_{n-2}$, **provided** the usage of $\mu_{n-1} = \frac{1}{2}\mu_{n-2}$ **does not endanger** the (unique) existence of the probability distribution of X (i.e. the probability distribution uniquely determined by the solution of the system of equations (4.4) on taking $m = n - 1$).*

Proof of the proposition 5.5.1. Let us examine the difference between the **upper** ($ub(\mu_n)$) and the **lower** ($lb(\mu_n)$) bounds of μ_n handled by the proven inequality (5.4), namely the expression

$$\begin{aligned} &\mu_{n-1} - \frac{\mu_{n-1}^2}{\mu_{n-2}} \\ &= \underbrace{\frac{\mu_{n-1}}{\mu_{n-2}}\left(1 - \frac{\mu_{n-1}}{\mu_{n-2}}\right)}_{\leq \frac{1}{4} \text{ for every } n \in \mathbb{N}\setminus\{1\}} \mu_{n-2} \\ &\leq \frac{1}{4}\mu_{n-2} \end{aligned} \tag{5.12}$$

This shows that, the **least upper bound** of $ub(\mu_n) - lb(\mu_n) = \mu_{n-1} - \frac{\mu_{n-1}^2}{\mu_{n-2}}$

with respect to all possible legitimate choices of μ_{n-1}, is nothing but $\frac{1}{4}\mu_{n-2}$ and this very least upper bound is **attained**, when $\mu_{n-1} = \frac{1}{2}\mu_{n-2}$.

Now, let us consider the two following points:

- If $\mu_{n-1} - \frac{\mu_{n-1}^2}{\mu_{n-1}} \leq \epsilon$ is **made to be fulfilled** for the **least possible** value of n, we have already denoted this very **minimum** value of n as $\mathbf{N_{optimal}}$.
- Again, if $\frac{1}{4}\mu_{n-2} \leq \epsilon$ is **made to be fulfilled** for the **least possible** value of n, let us denote this very **minimum** value by $\mathbf{N_{ultimate}}$, i.e.
  - $\frac{1}{4}\mu_{\mathbf{N_{ultimate}}-2} \leq \epsilon$
  - $\frac{1}{4}\mu_{\mathbf{N_{ultimate}}-3} > \epsilon$

Understandably, the **minimum** value of n needed to fulfill $\frac{1}{4}\mu_{n-2} \leq \epsilon$ is **greater** than (or **possibly equal** to) the **minimum** value of n needed to fulfill $\mu_{n-1} - \frac{\mu_{n-1}^2}{\mu_{n-1}} \leq \epsilon$.

That is, $\mathbf{N_{optimal}} - 1 \leq \mathbf{N_{ultimate}} - 1$, which means

$$\mathbf{N_{optimal}} \leq \mathbf{N_{ultimate}} \tag{5.13}$$

the **equality** of which holds, **if** $\mu_{n-1} = \frac{1}{2}\mu_{n-2}$ **legitimately** holds.

In that case, it is conclusively clear that the **worst possible value** of m shall be $\mathbf{N_{ultimate}} - 1$. This is precisely to say that, with reference to the statement 5.1.3, **any value** of m chosen that is higher than $\mathbf{N_{ultimate}} - 1$ is completely **unpracticable** or **unjustified**. □

Remark 5.5.1. *Notably, μ_n being the n^{th} moment of X, the **insignificance** of μ_n (X being discrete or continuous), **increases monotonically** with the **increase** in n.*

5.6 Existing difficulties in the discrete cases of X

In this section, we shall **briefly** discuss about the existing difficulties about the discussions of the **smallness of the bounding range** $ub(\mu_n) - lb(\mu_n)$ of the moment μ_n, in case X happens to be discrete.

In order to discuss the same, the smallness of $ub(\mu_n) - \mu_n$ and of $\mu_n - lb(\mu_n)$ shall be discussed separately. For this, the **expressions** of $ub(\mu_n) - \mu_n$ and $\mu_n - lb(\mu_n)$ are deduced.

Here, the **predetermined** support $\mathcal{X}_X = \{0 = x_1, x_2, \ldots, x_N = 1\}$ of the probability distribution of the discrete X is subjected to $0 = x_1 < x_2 < x_3 < \ldots < x_N = 1$.

Let us **reconsider** and **reestablish** both the inequalities $\mu_n < \mu_{n-1} = ub(\mu_n)$ (i.e. (5.9)) and $lb(\mu_n) = \frac{\mu_{n-1}^2}{\mu_{n-2}} < \mu_n$ (i.e. (5.7)) for the **discrete** X briefly one by one as follows:

5.6.1 The restatement of the inequality $\mu_n < \mu_{n-1}$ for a discrete X

The inequality $\mu_n < \mu_{n-1}$ for any $n \in N$ can be conveniently modified to the give the following proposition:

Proposition 5.6.1 (The upper bound of μ_{n+1} for any $n \in \mathbb{N}_0$). *The inequality reads $\mu_{n+1} < \mu_n = ub(\mu_{n+1})$.*

***Elaboration** of the **proposition 5.6.1**.*

$$ub(\mu_{n+1}) - \mu_{n+1} = \mu_n - \mu_{n+1} = \sum_{j=1}^{N} x_j^n (1 - x_j) e^{p_X(x_j)} \tag{5.14}$$

the polynomial $p_X(x)$ in x is being defined by (5.5).

Because of $0 = x_1 < x_2 < \ldots < x_{N-1} < x_N = 1$, the (5.14) can be rewritten as

$$\mu_n - \mu_{n+1} = \sum_{j=2}^{N-1} x_j^n (1 - x_j) e^{p_X(x_j)} > 0 \text{ for } n \in \mathbb{N}_0 \tag{5.15}$$

which implies nothing but $\mu_{n-1} > \mu_n$ for every $n \in \mathbb{N}$ and thereby the inequality (5.9) gets reestablished. □

5.6.2 The reestablishment of the inequality $\mu_n > \frac{\mu_{n-1}^2}{\mu_{n-2}}$ for a discrete X

The inequality $\mu_n > \frac{\mu_{n-1}^2}{\mu_{n-2}}$ for any $n \in N$ can be conveniently modified to the give the following proposition:

Proposition 5.6.2 (The lower bound of μ_{n+2} for any $n \in \mathbb{N}_0$). *The inequality reads* $\mu_{n+2} > \frac{\mu_{n+1}^2}{\mu_n} = lb(\mu_{n+2})$.

***Reestablishment** of the **proposition 5.6.2**.*

$$\mu_{n+2} - lb(\mu_{n+2}) = \mu_{n+2} - \frac{\mu_{n+1}^2}{\mu_n} \tag{5.16}$$

the polynomial $p_X(x)$ in x is being defined by (5.5).

Therefore, by using the (5.16), we get

$$\begin{aligned}
\mu_{n+2}\mu_n - \mu_{n+1}^2 &= \sum_{j=1}^{N} x_j^{n+2} e^{p_X(x_j)} \sum_{j=1}^{N} x_j^n e^{p_X(x_j)} - \left(\sum_{j=1}^{N} x_j^{n+1} e^{p_X(x_j)} \right)^2 \\
&= \sum_{j=1}^{N} x_j^{2n+2} \left(e^{p_X(x_j)} \right)^2 + \sum_{j<k} \left(x_j^{n+2} x_k^n + x_j^n x_k^{n+2} \right) e^{p_X(x_j)} e^{p_X(x_k)} \\
&\quad - \left(\sum_{j=1}^{N} x_j^{2n+2} \left(e^{p_X(x_j)} \right)^2 + 2 \sum_{j<k} x_j^{n+1} x_k^{n+1} e^{p_X(x_j)} e^{p_X(x_k)} \right) \\
&= \sum_{j<k} \Big(x_j^n x_k^n \left(x_j^2 + x_k^2 \right) e^{p_X(x_j)} e^{p_X(x_k)} \\
&\qquad\qquad - 2 x_j^n x_k^n \left(x_j x_k \right) e^{p_X(x_j)} e^{p_X(x_k)} \Big) \\
&= \sum_{j<k} \left(x_j^n x_k^n \left(x_j - x_k \right)^2 e^{p_X(x_j)} e^{p_X(x_k)} \right) > 0
\end{aligned} \tag{5.17}$$

the polynomial $p_X(x)$ in x is being defined by (5.5).

Obviously, (5.17) leads us to

$$\mu_{n+2} > \frac{\mu_{n+1}^2}{\mu_n} \tag{5.18}$$

which implies nothing but $\mu_n > \frac{\mu_{n-1}^2}{\mu_{n-2}}$ for every $n \in \mathbb{N} \setminus \{1\}$ and thereby the inequality (5.7) gets reestablished. □

5.6.3 General remarks

Remark 5.6.1. *Clearly, the judgement of the smallness of the finite sums* $\sum_{j=2}^{N-1} x_j^n(1-x_j)e^{p_X(x_j)} > 0$ *(given by 5.15) and* $\sum_{j<k} \left(x_j^n x_k^n \left(x_j - x_k\right)^2 e^{p_X(x_j)} e^{p_X(x_k)} \right) > 0$ *(given by 5.17) for judging the smallness of* $ub(\mu_{n+1}) - \mu_{n+1}$ *and* $\mu_{n+2} - lb(\mu_{n+2})$ *respectively is* ***certainly not simple****. The* ***principle problem*** *lies with the very fact that the smallness of both these expressions depends* ***additionally*** *on the predetermined values of* x_j*'s (*$j \in \{2, 3, \ldots, N-1\}$*) and obviously on* N*. This is* ***unlike*** *the continuous case of* X*.*

Remark 5.6.2. ***From the stochastic point of view or even from the programming point of view****, in discrete cases, as far as the determination of the* ***greatest lower bound*** *of the second moment* μ_2 *of* X *(or equivalently the same of the second moment* $\mu_Y^{(2)}$ *of* Y*) is concerned, a more careful investigation is necessary, because in certain cases (as we have already mentioned)* $lb(\mu_2) < glb(\mu_2)$ *holds,* ***unlike*** *in continuous cases of* X*. However, we shall see that, for the* ***least upper bound*** *of* μ_2*,* $ub(\mu_2) = lub(\mu_2)$ ***always*** *holds.*

So, for **discrete** cases of X (as we have already mentioned), let us confine our coming discussions about the **greatest lower bound** and the **least upper bound** of μ_n **to** $n = 2$ only and avoid the discussions about the same for $n \geq 3$.

5.7 Bounds of μ_2 for a discrete X

5.7.1 The $lub(\mu_2)$ for a discrete X

We arrive at the following proposition

Proposition 5.7.1 ($lub(\mu_2) = ub(\mu_2)$ holds for a discrete X). *Our proposition says that*

$$lub(\mu_2) - \mu_2 = \mu_1 - \mu_2 = \sum_{j=1}^{N} x_j(1-x_j)e^{p_X(x_j)} \tag{5.19}$$

can be made infinitesimally small *for any predetermined support $\mathcal{X}_X = \{0 = x_1, x_2, \ldots, x_N = 1\}$, where the polynomial $p_X(x)$ in x is defined by (5.5).*

*Proof of the **proposition 5.7.1**.* In order to reason the statement (5.19) saying that $\mathcal{X}_X$ **does not affect** the infinitesimal nature of $lub(\mu_2) - \mu_2$, we present the two following arguments:

- The finite sum (5.19) is basically equal to the following finite sum $\sum_{j=2}^{N-1} x_j(1-x_j)e^{p_X(x_j)}$, simply because of $x_1 = 0$ and $x_N = 1$.
- The second moment μ_2 of X can always be chosen **sufficiently close** to the first moment μ_1 of X **from left**, so that the sum of the two individual probabilities $e^{p_X(x_1)}$ and $e^{p_X(x_N)}$ on the extreme ends i.e. the sum $e^{p_X(x_1)} + e^{p_X(x_N)}$ can be made can be made **infinitesimally close** to 1 from left, with the restriction of $e^{p_X(x_1)} < 1$ and $e^{p_X(x_N)} < 1$. This necessarily means that the sum of the other individual probabilities, i.e. $\sum_{j=2}^{N-1} e^{p_X(x_j)} = 1 - (e^{p_X(x_1)} + e^{p_X(x_N)})$ can be made **infinitesimally small**. This means, because of $x_j(1-x_j) < 1$, the finite sum $\sum_{j=2}^{N-1} x_j(1-x_j)e^{p_X(x_j)} = \mu_1 - \mu_2$ **can be made infinitesimally small**

and thereby proving our **proposition (5.7.1)** confirming $lub(\mu_2) = ub(\mu_2)$. □

5.7.2 The $glb(\mu_2)$ for a discrete X

We arrive at the following proposition

Proposition 5.7.2 ($glb(\mu_2) \geq lb(\mu_2)$ holds for a discrete X). *Our proposition says that*

$$\mu_2 - lb(\mu_2) = \mu_2 - \mu_1^2 = \sum_{j<k} \Big((x_j - x_k)^2 \, e^{p_X(x_j)} e^{p_X(x_k)} \Big) \tag{5.20}$$

cannot be made infinitesimally small *for any predetermined support $\mathcal{X}_X = \{0 = x_1, x_2, \ldots, x_N = 1\}$, where the polynomial $p_X(x)$ in x is defined by (5.5).*

Proof of the ***proposition 5.7.2****.* For a discrete X, the predetermined $\mathcal{X}_X = \{0 = x_1, x_2, \ldots, x_N = 1\}$ plays a role. We shall show that the infinitesimal nature of $\mu_2 - glb(\mu_2)$ **is affected** by $\mathcal{X}_X$ and the existence of the probability distribution of X is possible with subject to the **fulfillment of a condition** imposed on $\mathcal{X}_X$.

Contrarily, in cases when X is continuous, then the infinitesimal nature of $\mu_2 - glb(\mu_2)$ **is not affected** by the predetermined $\mathcal{X}_X = [0, 1]$ at all.

Here, we intend to show that μ_1^2 can be the greatest lower bound of μ_2 for a discrete X, **only if** $\mathcal{X}_X$ fulfills a specific condition.

The basic content of our discussion would be about what exactly happens to the existence of probability distribution of X, if μ_2 is made to approach close to μ_1^2 **from the right**. We shall see that the existence of this probability distribution has basically to do with the **position of the first moment** μ_1 within $\mathcal{X}_X$, principally because μ_1 is regarded as the **center of mass** or the **position parameter** of any probability distribution of X for practical use.

For this, by using the inequality (5.17) for the special case of $n = 0$, we have

$$\begin{aligned} \mu_2\mu_0 - \mu_1^2 &= \sum_{j<k} \Big((x_j - x_k)^2 \, e^{p_X(x_j)} e^{p_X(x_k)} \Big) \\ &> (x_s - x_{s+1})^2 \, e^{p_X(x_s)} e^{p_X(x_{s+1})} > 0 \\ \Rightarrow \quad & \mu_2 > \mu_1^2 \end{aligned} \tag{5.21}$$

such that x_s and x_{s+1}, $s \in \{1, 2, \ldots, N-1\}$ are the two successive elements of $\mathcal{X}_X = \{0 = x_1, x_2, \ldots, x_N = 1\}$, within which μ_1 lies, i. e. $x_s \leq \mu_1 \leq x_{s+1}$. Moreover, the inequality (5.7) gets reconfirmed here in this case.

Trivially, for $\mu_0 = 1$, the inequality (5.21) gets rewritten as:

$$\begin{aligned} \mu_2 - \mu_1^2 &= \sum_{j<k} \Big((x_j - x_k)^2 \, e^{p_X(x_j)} e^{p_X(x_k)} \Big) \\ &> (x_s - x_{s+1})^2 \, e^{p_X(x_s)} e^{p_X(x_{s+1})} \end{aligned} \tag{5.22}$$

Therefore, if μ_2 is made to approach closer to μ_1^2 from the right, then the expression $(x_s - x_{s+1})^2 \, e^{p_X(x_s)} e^{p_X(x_{s+1})}$, which is a function of s, is obviously the **most deciding term** determining the **infinitesimal nature** of the aforesaid finite sum (5.21), namely

$$\sum_{j<k} \Big((x_j - x_k)^2 \, e^{p_X(x_j)} e^{p_X(x_k)} \Big)$$

simply because, due to the very fact that at least one of the two elements x_s and x_{s+1} (if not both) is the element **closest** to μ_1, at least one of the two individual probabilities $e^{p_X(x_s)}$ and $e^{p_X(x_{s+1})}$ (if not both) must be **significantly more important** than all the other $N-1$ (or $N-2$) probability elements belonging to the **discrete** probability distribution of X.

In fact, in that case, at least one of these two individual probabilities $e^{p_X(x_s)}$ and $e^{p_X(x_{s+1})}$ (if not both) must be **significantly larger** than all the other $N-1$ (or $N-2$) probability elements of the probability distribution of X.

Now, because of $0 < \mu_1 < 1$, the following two cases do arise:

- $$x_s < \mu_1 < x_{s+1} \text{ for a particular } s \in \{1, 2, \ldots, N-1\} \tag{5.23}$$

  Here, **both** the individual probabilities $e^{p_X(x_s)}$ and $e^{p_X(x_{s+1})}$ are **significantly** larger than all the other $N-2$ probabilities belonging to the probability distribution of X.

- $$\mu_1 = x_s \text{ for a particular } s \in \{2, 3, \ldots, N-1\} \tag{5.24}$$

  Here, the probability $e^{p_X(x_s)}$ is **significantly** larger than all the other $N-1$ probabilities belonging to the probability distribution of X.

Thus, if we have a close look at the inequality part of (5.22), it is well observable that, if μ_2 is to be chosen arbitrarily close to μ_1^2 from right, then

the expression $(x_s - x_{s+1})^2 e^{p_X(x_s)} e^{p_X(x_{s+1})}$ must be in a position to be made **arbitrarily** small.

Therefore, by setting $\Delta_s = x_{s+1} - x_s$, the arbitrary smallness of the expression $(x_s - x_{s+1})^2 e^{p_X(x_s)} e^{p_X(x_{s+1})}$ with regard to $\mu_2 \to \mu_1^2$ from right necessitates an intensive investigation:

This is the situation, where the predetermined mean (μ_1) and the predetermined second moment (μ_2) (or equivalently the variance ($\sigma^2 = \mu_2 - \mu_1^2$)) should determine the desired uni- extremal probability distribution of X.

In fact, as our current problem says, if μ_2 is made to be sufficiently close to its $lb(\mu_2) = \mu_1^2$, this is the case of an **uni-modal probability distribution** of X.

The discussion of this uni-modality must be divided into two following cases:

<u>**Case 1**</u>: $x_s < \mu_1 < x_{s+1}$ for a particular $s \in \{1, 2, \ldots, N-1\}$ (referred to (5.23)):

Here, we can well observe that **neither the probability** $e^{p_X(x_s)}$, **nor the probability** $e^{p_X(x_{s+1})}$ can be arbitrarily small and hence for any fixedly chosen $\Delta_s = x_{s+1} - x_s$, the expression $\Delta_s^2 e^{p_X(x_s)} e^{p_X(x_{s+1})}$ **cannot be arbitrarily small**.

That is, by (5.22), the expression $\mu_2 - \mu_1^2 = \sigma^2$ cannot be arbitrarily small, simply because $\mu_2 - \mu_1^2 > \Delta_s^2 e^{p_X(x_s)} e^{p_X(x_{s+1})} > 0$.

So, the problem demands that we must impose a condition on Δ_s. If this condition fails to be fulfilled, the probability distribution of X **cannot exist**. In other words, with subject to every predetermined $\mathcal{X}_X$ involving Δ_s and the predetermined moments μ_1, μ_2, the probability distribution of X will not exist, unless we give the following condition imposed on Δ_s:

For any given arbitrarily small $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\sigma^2 = \mu_2 - \mu_1^2 < \epsilon \text{ for every } \Delta_s < \delta \tag{5.25}$$

which means, for every choice of μ_2 with regard to the fulfillment of $\mu_2 > \mu_1^2$, the choice being described by ϵ, $\mathcal{X}_X$ must fulfill a **condition for the existence of the probability distribution of** X in form of $\Delta_s < \delta$.

Hence in this **case 1**, we have $glb(\mu_2) > lb(\mu_2)$.

In particular, if the elements of $\mathcal{X}_X$ are **equidistant**, then $\Delta_s = \frac{1}{N}$ and in that case, the **condition imposed on** $\mathcal{X}_X$ would be $\frac{1}{N} < \delta$.

<u>**Case 2**</u>: $\mu_1 = x_s$ for a particular $s \in \{2, 3, \ldots, N-1\}$ (referred to (5.24)):

Here, we can well observe that **the probability** $e^{p_X(x_s)}$ can be **arbitrarily close to 1** and **the probability** $e^{p_X(x_{s+1})}$ can be **arbitrarily close to 0**, i. e. arbitrarily small. Hence, for any fixedly chosen $\Delta_s = x_{s+1} - x_s$, the expression $\Delta_s^2 e^{p_X(x_s)} e^{p_X(x_{s+1})}$ **can be arbitrarily small unconditionally**.

Thus, in this case, for any arbitrary choice of μ_2 with subject to $\mu_2 > \mu_1^2$, the probability distribution of X **does always exist** with subject to the predetermined $\mathcal{X}_X$.

That is, **no condition needs to be imposed on** $\mathcal{X}_X$ in this case.

Hence in this **case 2**, we have $glb(\mu_2) = lb(\mu_2)$.

<u>**Conclusive statement**</u>: Hence, by the cases 1 and 2, our **proposition (5.7.2)** gets **proved** and thereby confirming $glb(\mu_2) \geq lb(\mu_2)$. □

5.8 Summary pertaining to the bounds of μ_2

To summarize everything regarding the bounds of μ_2, the inequality (5.4) gives the following:

- μ_1 is the least upper bound of μ_2, both in discrete and continuous cases, but
- μ_1^2 is the greatest lower bound of μ_2 in continuous cases unconditionally, but only with subject to the fulfillment of a condition, namely (5.25) imposed on $\mathcal{X}_X$ in discrete cases

Notably, since the discussion about the greatest lower bound of μ_2 for a discrete X is complicated enough, the discussions about the greatest lower bounds of μ_n, $n \geq 3$ for a discrete X are **even more complicated**.

5.9 Bounds of the first two moments of Y

For programming purposes, the upper bound and the lower bound of

- the first two moments of Y
- the second central moment of Y, namely of the variance of Y

are necessary guidelines addressed to the users, especially for the continuous cases.

5.9.1 Upper and lower bounds of the first and the second moment of Y

Let us consider the following well known inequalities once again, which are the upper bounds and lower bounds of the first two moments of the random variable X:

$$0 < E[X] < 1 \tag{5.26}$$

$$(E[X])^2 < E[X^2] < E[X] \tag{5.27}$$

For μ_1 and μ_2 being $E[X]$ and $E[X^2]$ respectively, the above relations (5.26) and (5.27) can also be written as

$$0 < \mu_1 < 1 \tag{5.28}$$

$$\mu_1^2 < \mu_2 < \mu_1 \tag{5.29}$$

and as a result, for $X = \frac{Y-a}{b-a}$, we have

$$0 < \frac{E[Y]-a}{b-a} < 1 \text{ and} \tag{5.30}$$

$$\left(\frac{E[Y]-a}{b-a}\right)^2 < \frac{E\left[Y^2\right]-2aE[Y]+a^2}{(b-a)^2} < \frac{E[Y]-a}{b-a} \tag{5.31}$$

which concludes to the following relations, which are the upper the lower bound of the first two moments of the random variable Y:

$$a \quad < E[Y] < b \tag{5.32}$$

$$(E[Y])^2 \quad < E[Y^2] < \Big(a+b\Big)E[Y] - ab \tag{5.33}$$

Again, for $\mu_Y^{(1)}$ and $\mu_Y^{(2)}$ being $E[Y]$ and $E[Y^2]$ respectively, the above inequalities can also be written as:

$$a \quad < \mu_Y^{(1)} < b \tag{5.34}$$

$$\Big(\mu_Y^{(1)}\Big)^2 \quad < \mu_Y^{(2)} < \Big(a+b\Big)\mu_Y^{(1)} - ab \tag{5.35}$$

Hence, (5.34) and (5.35) give the upper bound and the lower bound of the first and the second moment of $Y|\{d_Y\}$ respectively.

Important remark: As a matter of fact, although $\mu_2 < \sqrt{\mu_3\mu_1} < \mu_1$, the least upper bound of μ_2 is μ_1. This is because any upper bound of μ_2 has to consist of moment(s) of order less than 2. This explains why we could term this l.u.b as the conditional l.u.b.

5.9.2 Upper and the lower bound of the second central moment of Y

By taking $Var[X] = E\left[X^2\right] - (E[X])^2$, we get the l.u.b. and g.l.b. of $Var[X]$ as

$$0 < Var[X] < E[X]\Big(1 - E[X]\Big) \tag{5.36}$$

which leads us to

$$0 < Var\left[\frac{Y-a}{b-a}\right] < \left(\frac{E[Y]-a}{b-a}\right)\left(\frac{b-E[Y]}{b-a}\right) \tag{5.37}$$

and therefore

$$0 < Var[Y] < \Big(E[Y]-a\Big)\Big(b-E[Y]\Big) \tag{5.38}$$

For σ_Y^2 to be the $Var[Y]$, we get

$$0 < \sigma_Y^2 < \left(\mu_Y^{(1)} - a\right)\left(b - \mu_Y^{(1)}\right) \tag{5.39}$$

Hence, (5.39) gives the upper bound and the lower bound of the second central moment of Y.

For our future references, if σ^2 be the $Var[X]$, we shall make use the following relation

$$\sigma = \frac{\sigma_Y}{b-a} \tag{5.40}$$

5.10 A preliminary note on the random variable X

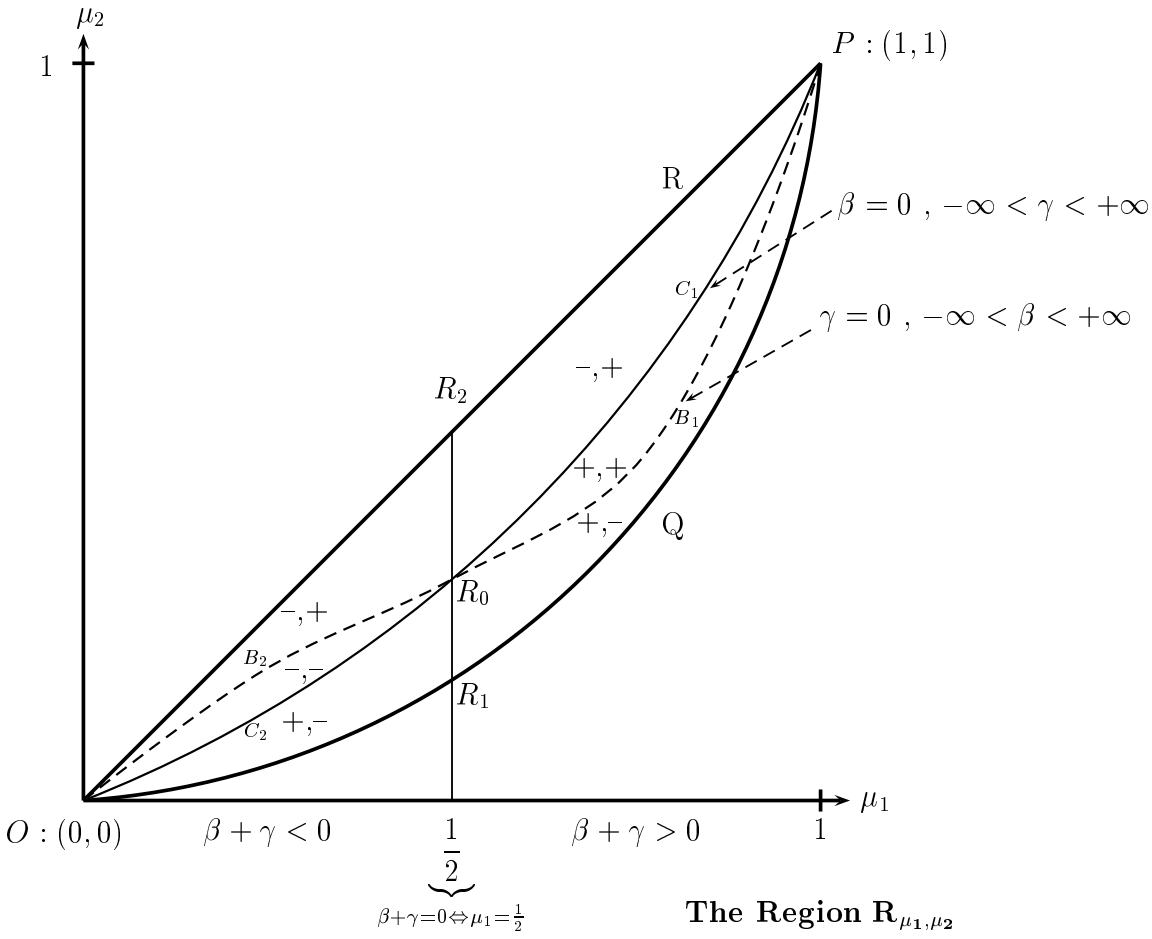
Before we give elaborated discussions about the standard minimum information probability distributions of the random variable Y (or equivalently of the random variable $X = \frac{Y-y_1}{y_N-y_1}$ in the **discrete** case or $X = \frac{Y-a}{b-a}$ in the **continuous** case) in the very next chapter, let us state briefly the discuss about **the most frequently used moments** of X preliminarily.

The probability distributions of monotonic type and of uni-extremal type are principally discussed in this thesis and therefore, in accordance with the minimum information principle, the construction of the following probability distribution types:

Definition 5.10.1 (Recapitulation of a monotone probability distribution type). *A **monotonic** probability distribution of X necessitates, **in addition to** the support $\mathcal{X}_X$, the knowledge of the **first** moment only, where* $\mu_1 = \frac{\int_{\mathcal{X}_X} xe^{\beta x}\nu_X(dx)}{\int_{\mathcal{X}_X} e^{\beta x}\nu_X(dx)}$ *. The first moment μ_1 is restricted within the region $R_{\mu_1} = \{\mu_1 \mid 0 < \mu_1 < 1\}$. The relationship between β and μ_1 is of extreme importance and is rather simple.*

Definition 5.10.2 (Recapitulation of a uni-extremal probability distribution type). *A **uni-extremal** probability distribution of X necessitates, **in addition to** the support $\mathcal{X}_X$, the knowledge of the **first** and the **second** moment only, where* $\mu_k = \frac{\int_{\mathcal{X}_X} x^k e^{\beta x+\gamma x^2}\nu_X(dx)}{\int_{\mathcal{X}_X} e^{\beta x+\gamma x^2}\nu_X(dx)}$, $k \in \{1, 2\}$ *.*
The first moment μ_1 and the second moment μ_2 are restricted within the region $R_{\mu_1,\mu_2} = \{(\mu_1, \mu_2) \mid 0 < \mu_1 < 1, \mu_1^2 < \mu_2 < \mu_1\}$. The relationship between (β, γ) and (μ_1, μ_2) is (especially, when X is continuous, the picture is more or less similar, when X is discrete) of extreme importance and is presented graphically[4] in the very next page.

[4]This graphical representation of the region R_{μ_1,μ_2} is basically needed for the discussions of the subsection 6.3.10. Because of technical difficulties, this graphical representation is given in this very chapter.



The Region $\mathbf{R}_{\mu_1,\mu_2}$

Chapter 6

Characteristic properties of standard m. i. probability distributions

For the sake of simplicity, **without any loss of generality**, we shall confine our discussions to the probability distributions of X only and not the same of Y in this chapter.

The discussions regarding the probability distributions of Y would mean a very simply linear transformation, namely $Y = a + (b - a)X$ and that should not be an issue for intensive considerations at all.

We shall therefore discuss about the characteristic properties of standard minimum information probability distributions of the random variable X. As we know, standard minimum information probability distribution may be a constant, monotone or an uni-extremal one. The constant cases are too trivial and therefore the discussions in this regard are omitted here. In other words, we shall analyze the characters of the standard minimum probability distributions of X.

6.1 Preliminaries

6.1.1 The basic idea

In the system of equations (4.4), by setting $\beta_1 = \beta$, $\beta_2 = \gamma$ and $\beta_i = 0$ for every $i \geq 2$, we shall write μ_i ($i \in \{1, 2\}$) in form of the following statement in the following way:

Statement 6.1.1 (Recapitulation of μ_1 and μ_2).

$$\mu_i = \frac{\int\limits_{\mathcal{X}_X} x^i e^{\beta x + \gamma x^2} \nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta x + \gamma x^2} \nu_X(dx)} = \begin{cases} \dfrac{\sum\limits_{j=1}^{N} x_j^i e^{\beta x_j + \gamma x_j^2}}{\sum\limits_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} & : \quad X \text{ is discrete} \\ \\ \dfrac{\int\limits_0^1 x^i e^{\beta x + \gamma x^2} dx}{\int\limits_0^1 e^{\beta x + \gamma x^2} dx} & : \quad X \text{ is continuous} \end{cases} \tag{6.1}$$

such that

- $\mathcal{X}_X = \{0 = x_1, x_2, \ldots, x_N = 1\}$, *if* X *is discrete and*
- $\mathcal{X}_X = [0, 1]$, *if* X *is continuous*

As we have already mentioned, we shall go by the two following statements:

Statement 6.1.2. *The discussions of the* ***minimum information monotone probability distributions*** *are subjected to the usage of* μ_1 *only, i.e.* $i = 1$ *only*

Statement 6.1.3. *The discussions of the* ***minimum information uniextremal probability distributions*** *are subjected to the usage of two moments* μ_1 *and* μ_2 *only, i.e.* $i \in \{1, 2\}$ *only*

6.1.2 The fulfillment's plausibility check

For the sake of a simple plausibility check, we shall show that the **necessary** and **sufficient** condition given in the subsection 4.4.2 for the existence of the probability distribution of X determined by the solution of the system of equations (4.4) is **fulfilled** in cases for $m \in \{1, 2\}$.

The fulfillment of the aforesaid necessary and sufficient condition in cases for $m \in \{1, 2\}$ necessitates the usage of $0 < \mu_1 < 1$ and $\mu_1^2 < \mu_2 < \mu_1$ (i.e. because of $\sigma^2 = \mu_2 - \mu_1^2 > 0$), we arrive at the two following propositions:

Proposition 6.1.1 (Plausibility check for monotone cases). *In **monotone cases**, the given **finite sequence** of moments, namely the set $\{\mu_0, \mu_1\}$ of moments, which is involved in finding the value of β for the determination of each of the monotonic probability distributions of X in both **discrete** and **continuous** cases, **fulfills** the condition for the existence of the solution of (6.2)*

*Establishing the **proposition 6.1.1**.* Here,

- by (4.24), $\det(\underline{H}_1) = \mu_1 > 0$
- by (4.31), $\det(\overline{H}_1) = 1 - \mu_1 > 0$

which gives $0 < \mu_1 < 1$ □

Proposition 6.1.2 (Plausibility check for uni-extremal cases). *In **uni-extremal cases**, the given **finite sequence** of moments, namely the set $\{\mu_0, \mu_1, \mu_2\}$ of moments, which is involved in finding the values of β and γ for the determination of each of the uni- extremal probability distributions of X in **continuous** cases, **fulfills** the condition for the existence of the solution of (6.8)*

Establishing the proposition 6.1.2. Here,

- by (4.21), $\det(\underline{H}_2) = \begin{vmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{vmatrix} = \mu_2 - \mu_1^2 > 0$
- by (4.28), $\det(\overline{H}_2) = \mu_1 - \mu_2 > 0$

which gives $\mu_1^2 < \mu_2 < \mu_1$ in **additional to** the proven $0 < \mu_1 < 1$.

However, if X happens to be **discrete**, then the existence of the probability distribution of X is not possible for **every** predetermined $\mathcal{X}_X$ consisting of a finite number of elements, despite fulfillment of the aforesaid necessary and sufficient condition **involving moments only** (i.e. **not involving** $\mathcal{X}_X$). □

Importantly, for our analysis, we shall have to give the **formal proof of the existence of the (unique) solution of the system of equations (4.2) (or equivalently of (4.4)) in cases for** $m \in \{1, 2\}$.

Notably, the uniqueness of the solutions of both (6.2) and (6.8) have already been established.

Obviously, as we have already mentioned, for our convenience, we shall use the system (4.4) instead of (4.2) **equivalently**.

Our analysis necessitates the introductions of certain important lemmas, which shall be duly presented.

6.1.3 Dealing with the special cases

Now, for the **sake of the completeness of our analysis**, the discussions about the nature of the probability distributions of X corresponding to the following is a must:

- $0 \leq \mu_1 \leq 1$ for **monotone cases**, where the special cases are referred to $\mu_1 = 0$ and $\mu_1 = 1$.
- $0 \leq \mu_1 \leq 1$, $\mu_1^2 \leq \mu_2 \leq \mu_1$ for **uni-extremal cases**, where the special cases are referred to
  1. $\mu_2 = \mu_1^2$, $0 < \mu_1 < 1$
  2. $\mu_2 = \mu_1$, $0 < \mu_1 < 1$
  3. $\mu_2 = \mu_1 = 0$
  4. $\mu_2 = \mu_1 = 1$

Each of these above stated special cases representing **boundaries of the moment spaces $\mathbf{D}^1$** and $\mathbf{D}^2$ respectively (denoted by $\partial\mathbf{D}^1$ and $\partial\mathbf{D}^2$ respectively). These boundaries $\partial\mathbf{D}^1$ and $\partial\mathbf{D}^2$ represent two special forms of probability distributions of X, namely **degenerated probability distributions** or **Bernoulli probability distributions**. So, the probability mass function $f_{X|\{d\}}(x)$ of X describing the probability distribution of X has one of the following **four** supports:

1. $\mathcal{X}_X = \{0\}$
2. $\mathcal{X}_X = \{1\}$
3. $\mathcal{X}_X = \{\mu_1\}$
4. $\mathcal{X}_X = \{0, 1\}$

Therefore, by keeping these things in mind, we proceed as the subsequent subsections follow.

6.2 M. i. monotone probability distributions

If the random variable $Y|\{d_Y\}$ follows a monotonic probability distribution, such that $d_Y = (\mu_Y^{(1)})$ or equivalently $d = (\mu_1)$, then the computation of $\lambda(d_Y) = (\lambda_1)$ necessitates the solution of the following equation

$$\mu_1 = \frac{\int\limits_{\mathcal{X}_X} xe^{\beta x}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta x}\nu_X(dx)} \tag{6.2}$$

whose solution is unique, provided $N \geq 2$ holds in discrete cases. Here, the first moment $E[X]$, namely μ_1, is expressed as a function of β.

6.2.1 Lemma for the monotonicity

Proposition 6.2.1 (The derivative of μ_1 with respect to β). *It is absolutely clear by (6.2) that*

$$\begin{aligned} \frac{d\mu_1}{d\beta} &= \frac{1}{\left(\int\limits_{\mathcal{X}_X} e^{\beta x}\nu_X(dx)\right)^2}\left[\int\limits_{\mathcal{X}_X} x^2e^{\beta x}\nu_X(dx)\int\limits_{\mathcal{X}_X} e^{\beta x}\nu_X(dx) - \left(\int\limits_{\mathcal{X}_X} xe^{\beta x}\nu_X(dx)\right)^2\right] \\ &= \mu_2 - \mu_1^2 = \sigma^2 > 0 \end{aligned} \tag{6.3}$$

which evidently clarifies the very fact that μ_1 is a strictly monotonically increasing function of β, even in discrete cases for $N \geq 2$. This is simply because, a discrete case corresponding to $N = 1$ necessarily means that the probability distribution is degenerated, for which $\sigma^2 = 0$.

Remark 6.2.1. *The **interchangeability** of the operations $\int\limits_{\mathcal{X}_X}$ and $\frac{\partial}{\partial\beta}$ have already been discussed previously.*

6.2.2 The restatement subjecting to the uniqueness

Remark 6.2.2 (Uniqueness of the solution of the system (6.2)). *The strict positiveness of $\frac{d\mu_1}{d\beta}$ (referred to (6.3)), (i.e. the strict monotonicity of μ_1 with respect to β) leads to the very fact that, for any fixed value of μ_1, there can exist only one value of β.*

Hence, the solution of the equation (6.2) is unique. □

Remark 6.2.3. ***Importantly****, this* ***aforesaid*** *uniqueness of the solution of (6.2) is tantamount to the uniqueness of the solution of each of the equations (11.17) and (11.24) belonging to monotone cases, together with the uniqueness of the solution of (4.2) for* $m = 1$. □

6.2.3 The existence of the solution of equation-system for $m = 1$

Theorem 6.2.1 (Existence for $m = 1$**).** *For the sake of convenience, let us take* X *for a* ***continuous*** *random variable* ***at first****. In that case, we get (6.2) as*

$$\mu_1 = \frac{\int\limits_0^1 xe^{\beta x}dx}{\int\limits_0^1 e^{\beta x}dx} = 1 + \frac{1}{e^{\beta}-1} - \frac{1}{\beta} \tag{6.4}$$

Then, the solution of (6.4) ***exists*** *and the* ***same*** *is the case for a* ***discrete*** X *as well.*

Proof of the ***theorem 6.2.1****.* Right now, by taking $\beta \to \infty$, we get by using L' Hospital's rule

$$\begin{aligned}
\lim_{\beta\to\infty}\left\{1 + \frac{1}{e^{\beta}-1} - \frac{1}{\beta}\right\} &= \lim_{\beta\to\infty}\frac{\beta e^{\beta} - e^{\beta} + 1}{\beta(e^{\beta}-1)}\left(\frac{\infty}{\infty}\right) \\
&= \lim_{\beta\to\infty}\frac{e^{\beta} + \beta e^{\beta} - e^{\beta}}{e^{\beta} + \beta e^{\beta} - 1} \\
&= \lim_{\beta\to\infty}\frac{\beta e^{\beta}}{e^{\beta} + \beta e^{\beta} - 1}\left(\frac{\infty}{\infty}\right) \\
&= \lim_{\beta\to\infty}\frac{e^{\beta} + \beta e^{\beta}}{e^{\beta} + e^{\beta} + \beta e^{\beta}} \\
&= \lim_{\beta\to\infty}\frac{\frac{1}{\beta} + 1}{\frac{2}{\beta} + 1} = 1
\end{aligned} \tag{6.5}$$

Again, by taking $\beta \to -\infty$, we get

$$\begin{aligned}
\lim_{\beta\to-\infty}\left\{1 + \frac{1}{e^{\beta}-1} - \frac{1}{\beta}\right\} &= \lim_{\beta\to-\infty}\left\{\frac{e^{\beta}}{e^{\beta}-1} - \frac{1}{\beta}\right\} \\
&= \frac{0}{0-1} - 0 = 0
\end{aligned} \tag{6.6}$$

Thus, by (6.6) and (6.5), it is evidently clear that, if β is made to run from $-\infty$ to ∞, then μ_1 runs from 0 to 1. In other words, by putting $\mu_1^{(\beta)} = \frac{\int_0^1 xe^{\beta x}dx}{\int_0^1 e^{\beta x}dx}$, the **co-domain of the function** $\mu_1^{(\beta)}$ (of β) **spans** the entire region $R_{\mu_1} = \{\mu \mid 0 < \mu_1 < 1\}$ **for all** $\beta \in \mathbb{R}$.

In other words (i.e. conversely), for every $\mu_1 \in R_{\mu_1}$ there exists an unique $\beta \in \mathbb{R}$. This proves the **existence** of the solution of (6.4) (or **equivalently** of (11.24)).

Now, let us draw our attention to the discrete case, i.e. let now us take X for a **discrete** random variable. In this case, the proof of the aforesaid existence is rather simple and can be given by considering a **basic characteristic property** of the probability mass function of X, viz. $f_{X|\{d\}}(x_j) = \frac{e^{\beta x_j}}{\sum_{j=1}^{N} e^{\beta x_j}}$, $j \in \{1, 2, \ldots, N\}$ for $0 = x_1 < x_2 < \ldots < x_N = 1$ in a very simpler manner. This characteristic property is described as

- If $\mu_1 \to 0 \Leftrightarrow \beta \to -\infty$, then the probability element $f_{X|\{d\}}(x_1)$ tends infinitesimally closer to 1 and all other probability elements, namely $f_{X|\{d\}}(x_2), f_{X|\{d\}}(x_3), \ldots, f_{X|\{d\}}(x_N)$ tend infinitesimally closer to 0.

- If $\mu_1 \to 1 \Leftrightarrow \beta \to \infty$, then the probability element $f_{X|\{d\}}(x_N)$ tends infinitesimally closer to 1 and all other probability elements, namely $f_{X|\{d\}}(x_1), f_{X|\{d\}}(x_2), \ldots, f_{X|\{d\}}(x_{N-1})$ tend infinitesimally closer to 0.

So, even in the discrete case, it is evidently clear that, all the possible real values of β (i.e. $\beta \in \mathbb{R}$) **span** the entire region $R_{\mu_1} = \{\mu \mid 0 < \mu_1 < 1\}$.

In other words (i.e. conversely), for every $\mu_1 \in R_{\mu_1}$ there exists an unique $\beta \in \mathbb{R}$. This proves the **existence** of the solution of (11.18) (or **equivalently** of (11.17)).

Whence, by combining both the discrete and the continuous cases of X, we conclude that the solution of (6.2) **exists** for every $\mu_1 \in R_{\mu_1}$.

This ultimately proves the **existence** of the solution of the system of equations (4.2) for $m = 1$. □

6.2.4 Characteristics of the density curves

As already shown with the help of (6.3), μ_1 increases strictly with the increase in β. In this regard, in the continuous case of X, by (6.4), the following conclusion can be easily drawn:

$$\begin{aligned} \mu_1 < \frac{1}{2} &\Leftrightarrow \beta < 0 \\ \mu_1 = \frac{1}{2} &\Leftrightarrow \beta = 0 \\ \mu_1 > \frac{1}{2} &\Leftrightarrow \beta > 0 \end{aligned} \tag{6.7}$$

This aforesaid conclusion (6.7) shall be **used to prove the existence of the solution of the system of equations (6.8)**.

By (6.7), it is evidently clear that the probability density curve $f_{X|\{d\}}(x) = \frac{e^{\beta x}}{\int_0^1 e^{\beta x} dx}$, $0 \leq x \leq 1$ has the following **characteristics**:

- it is strictly monotone increasing, if $\mu_1 > \frac{1}{2}$
- it is strictly monotone decreasing, if $\mu_1 < \frac{1}{2}$
- it is a constant curve, if $\mu_1 = \frac{1}{2}$ and thereby representing a constant probability distribution.

6.2.5 Probability distributions represented by boundary points defined by $\mu_1 \in \partial \mathbf{D}^1$

Let us discuss the special cases referred to the monotonic probability distributions of X one by one

1. **Case for** $\mu_1 = 1 \Leftrightarrow \beta = +\infty$:

   $f_{X|\{d\}}(x)$ can be defined to represent a **discrete degenerated probability distribution** defined by $f_{X|\{d\}}(x) = 1$ for $x = 1$

2. **Case for** $\mu_1 = 0 \Leftrightarrow \beta = -\infty$:

   $f_{X|\{d\}}(x)$ can be defined to represent a **discrete degenerated probability distribution** defined by $f_{X|\{d\}}(x) = 1$ for $x = 0$

This **fulfills the completeness** of our analysis of monotonic minimum information probability distributions.

6.3 M. i. uni-extremal probability distributions

If the random variable $Y|\{d_Y\}$ follows an uni-extremal probability distribution, such that $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$ or equivalently $d = (\mu_1, \mu_2)$, then the computation of $\lambda(d_Y) = (\lambda_1, \lambda_2)$ necessitates the solution of the following system of two simultaneous equations

$$\begin{cases} \mu_1 &= \dfrac{\int\limits_{\mathcal{X}_X} xe^{\beta x+\gamma x^2}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta x+\gamma x^2}\nu_X(dx)} \\ \\ \mu_2 &= \dfrac{\int\limits_{\mathcal{X}_X} x^2e^{\beta x+\gamma x^2}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta x+\gamma x^2}\nu_X(dx)} \end{cases} \tag{6.8}$$

whose solution is unique, provided $N \geq 3$ holds in discrete cases. Here, each of the first two moments $E[X]$ and $E[X^2]$, namely μ_1 and μ_2 respectively, are expressed as functions of β and γ.

6.3.1 First lemma for the uni-extremity

This lemma **enlists** the the partial derivatives of μ_1 and μ_2 with respect to β and γ respectively by using the system of equations (6.8). Additionally, the positivity of each of these partial derivatives shall be briefly and well established.

Here, the list of the statements are given as follows:

Statement 6.3.1 (The partial derivative $\frac{\partial \mu_1}{\partial \beta}$).

$$\begin{aligned} \frac{\partial \mu_1}{\partial \beta} &= \frac{\int\limits_{\mathcal{X}_X} x^2e^{\beta x+\gamma x^2}\nu_X(dx) \int\limits_{\mathcal{X}_X} e^{\beta x+\gamma x^2}\nu_X(dx) - \left(\int\limits_{\mathcal{X}_X} xe^{\beta x+\gamma x^2}\nu_X(dx)\right)^2}{\left(\int\limits_{\mathcal{X}_X} e^{\beta x+\gamma x^2}\nu_X(dx)\right)^2} \\ &= \mu_2 - \mu_1^2 = Var[X] > 0 \end{aligned} \tag{6.9}$$

Statement 6.3.2 (The partial derivative $\frac{\partial \mu_2}{\partial \gamma}$).

$$\begin{aligned}\frac{\partial \mu_2}{\partial \gamma} &= \frac{\int\limits_{\mathcal{X}_X} x^4 e^{\beta x+\gamma x^2}\nu_X(dx) \int\limits_{\mathcal{X}_X} e^{\beta x+\gamma x^2}\nu_X(dx) - \left(\int\limits_{\mathcal{X}_X} x^2 e^{\beta x+\gamma x^2}\nu_X(dx)\right)^2}{\left(\int\limits_{\mathcal{X}_X} e^{\beta x+\gamma x^2}\nu_X(dx)\right)^2} \\ &= \mu_4 - \mu_2^2 = Var[X^2] > 0\end{aligned} \tag{6.10}$$

and

Statement 6.3.3 (The partial derivative $\frac{\partial \mu_1}{\partial \gamma} = \frac{\partial \mu_2}{\partial \beta}$).

$$\begin{aligned}\frac{\partial \mu_1}{\partial \gamma} &= \frac{\partial \mu_2}{\partial \beta} \\ &= \frac{\int\limits_{\mathcal{X}_X} x^3 e^{\beta x+\gamma x^2}\nu_X(dx) \int\limits_{\mathcal{X}_X} e^{\beta x+\gamma x^2}\nu_X(dx) - \int\limits_{\mathcal{X}_X} x e^{\beta x+\gamma x^2}\nu_X(dx) \int\limits_{\mathcal{X}_X} x^2 e^{\beta x+\gamma x^2}\nu_X(dx)}{\left(\int\limits_{\mathcal{X}_X} e^{\beta x+\gamma x^2}\nu_X(dx)\right)^2} \\ &= \mu_3 - \mu_1\mu_2 \\ &> \frac{\mu_2^2}{\mu_1} - \mu_1\mu_2 \text{ (by using the inequality (5.4))} \\ &= \frac{\mu_2}{\mu_1}(\mu_2 - \mu_1^2) = \frac{\mu_2}{\mu_1} Var[X] > 0\end{aligned} \tag{6.11}$$

Remark 6.3.1. *The* ***interchangeability*** *of the operations* $\int\limits_{\mathcal{X}_X}$ *and* $\frac{\partial}{\partial \beta}$ *(or* $\frac{\partial}{\partial \gamma}$ *) have already been discussed previously.*

6.3.2 Second lemma for the uni-extremity

This lemma shall state and prove the following proposition:

Proposition 6.3.1 ($\frac{d\mu_2}{d\gamma}\Big|_{\mu_1=\mu_1^*} > 0$). *For any arbitrarily fixedly chosen value of* μ_1*, say* μ_1^**, such that* $0 < \mu_1^* < 1$*, the inequality* $\frac{d\mu_2}{d\gamma}\Big|_{\mu_1=\mu_1^*} > 0$ *holds.*

Proof of the ***proposition 6.3.1****.* We shall start by taking the differentials of μ_1 and μ_2 by using the system of simultaneous equations (6.8) at first, which is given as follows:

$$\begin{cases} d\mu_1 &= \frac{\partial\mu_1}{\partial\beta}d\beta + \frac{\partial\mu_1}{\partial\gamma}d\gamma \\ d\mu_2 &= \frac{\partial\mu_2}{\partial\beta}d\beta + \frac{\partial\mu_2}{\partial\gamma}d\gamma \end{cases} \tag{6.12}$$

For this, let the value of μ_1 be kept fixed, say $\mu_1 = \mu_1^*$ with regard to $0 < \mu_1^* < 1$ and μ_2 be allowed to vary within the range $\mu_1^{*2} < \mu_2 < \mu_1^*$.

Therefore, with subject to $\mu_1 = \mu_1^* = constant$ (i.e. $d\mu_1 = 0$), the system (6.12) takes the form

$$\begin{cases} 0 &= \frac{\partial\mu_1}{\partial\beta}d\beta + \frac{\partial\mu_1}{\partial\gamma}d\gamma \\ d\mu_2 &= \frac{\partial\mu_2}{\partial\beta}d\beta + \frac{\partial\mu_2}{\partial\gamma}d\gamma \end{cases} \tag{6.13}$$

from which, by (6.9), (6.10) and (6.11), we get

$$\begin{cases} 0 &= (\mu_2 - \mu_1^{*2})d\beta + (\mu_3 - \mu_1^*\mu_2)d\gamma \\ d\mu_2 &= (\mu_3 - \mu_1^*\mu_2)d\beta + (\mu_4 - \mu_2^2)d\gamma \end{cases} \tag{6.14}$$

which gives

$$\frac{d\beta}{d\gamma} = -\frac{\mu_3 - \mu_1^*\mu_2}{\mu_2 - \mu_1^{*2}} \tag{6.15}$$

and consequently

$$\begin{aligned} \frac{d\mu_2}{d\gamma} &= (\mu_3 - \mu_1^*\mu_2)\frac{d\beta}{d\gamma} + (\mu_4 - \mu_2^2) = (\mu_4 - \mu_2^2) - \frac{(\mu_3 - \mu_1^*\mu_2)^2}{\mu_2 - \mu_1^{*2}} \\ &= (\mu_4 - \mu_2^2)\left(1 - \rho_{X,X^2}^2\right)\Big|_{\mu_1=\mu_1^*} = \sigma_{X^2}^2\left(1 - \rho_{X,X^2}^2\right)\Big|_{\mu_1=\mu_1^*} > 0 \end{aligned} \tag{6.16}$$

simply because of the following:

1. The variance of the random variable X^2, namely $\sigma^2_{X^2} = \mu_4 - \mu_2^2$, is ought to be positive, because the probability distribution of X or of X^2 is not degenerated.

2. ρ_{X,X^2} being the correlation coefficient between the random variables X and X^2, we must necessarily have $\rho^2_{X,X^2} \leq 1$.

   Now, X^2 could be linear function of X, if $\mathcal{X}_X(\{d\})$ would have at most two elements, implying $N \leq 2$, which would necessarily mean $\rho^2_{X,X^2} = 1$. But, since we are given with $N \geq 3$, we must conclude that X^2 is not a linear function of X and therefore

$$\rho^2_{X,X^2} < 1 \tag{6.17}$$

   and consequently by

$$\rho^2_{X,X^2} = \frac{\{E\left[(X-\mu_1)(X^2-\mu_2)\right]\}^2}{Var[X]\ Var[X^2]} < 1 \tag{6.18}$$

   we got

$$\begin{aligned} &\left\{E\left[(X-\mu_1)(X^2-\mu_2)\right]\right\}^2 < Var[X]\ Var[X^2] \\ \Leftrightarrow &\left\{E\left[X^3 - X\mu_2 - X^2\mu_1 + \mu_1\mu_2\right]\right\}^2 \\ &\qquad < \left(E[X^2] - (E[X])^2\right)\left(E[X^4] - (E[X^2])^2\right) \\ \Leftrightarrow &\left\{\mu_3 - \mu_1\mu_2 - \mu_1\mu_2 + \mu_1\mu_2\right\}^2 < (\mu_2 - \mu_1^2)(\mu_4 - \mu_2^2) \\ \Leftrightarrow \quad &1 - \frac{(\mu_3 - \mu_1\mu_2)^2}{(\mu_4 - \mu_2^2)(\mu_2 - \mu_1^2)} = 1 - \rho^2_{X,X^2} > 0 \end{aligned} \tag{6.19}$$

Thus, we have established the following targeted inequality:

$$\left.\frac{d\mu_2}{d\gamma}\right|_{\mu_1=\mu_1^*} > 0 \tag{6.20}$$

and thereby establishing the **proposition 6.3.1**. □

6.3.3 Third lemma for the uni-extremity

The uniqueness of the solution of the system of simultaneous equations (6.8) follows immediately from the inequality (6.20) (i.e. directly from the established **proposition 6.3.1**) and this can be **reestablished** in two simple steps. We shall perform these steps in the following proposition:

Proposition 6.3.2 (Uniqueness of the solution of the system (6.8)). *The solution of system of equations (6.8) is* ***unique****.*

Proof of the ***proposition (6.3.2)****.* **First Step**: In the first step, we shall show that for any arbitrarily fixed values of μ_1 and μ_2, say μ_1^* and μ_2^* respectively with regard to $0 < \mu_1^* < 1$ and $\mu_1^{*2} < \mu_2^* < \mu_1^*$, the value of γ can be determined uniquely, say $\gamma = \gamma^*$.

Here, since μ_1^* has been an arbitrarily chosen value of μ_1 within the interval range $(0, 1)$, the inequality (6.20) leads us to the following:

$$\begin{gathered}\mu_2 \text{ is a strictly monotonically increasing function of } \gamma, \text{ with subject to} \\ \text{an arbitrarily fixed } \mu_1 = \mu_1^* \text{ within } (0, 1)\end{gathered} \tag{6.21}$$

This is to say that, for any arbitrarily fixed $\mu_1 = \mu_1^*$ within $(0, 1)$ and for any given $\mu_2 = \mu_2^* \in (\mu_1^{*2}, \mu_1^*)$, there can exist only an unique value of γ, say $\gamma = \gamma^*$.

In other words, for any two fixed values of μ_1 and μ_2, namely μ_1^* and μ_2^* respectively, there can exist only an unique value of γ, namely γ^*.

Second Step: In the second and the final step, we shall show that the value of β contained in the system (6.8) is unique too. Immediately after this, the targeted uniqueness follows conclusively.

Here, by (6.9), we have $\frac{\partial \mu_1}{\partial \beta} = \mu_2 - \mu_1^2 = \sigma^2 > 0$, which necessarily means that for the chosen fixed value of γ, namely γ^* and for the fixed value of μ_1, namely μ_1^*, the following equation is solvable uniquely for β:

$$\mu_1^* = \frac{\int\limits_{\mathcal{X}_X} x e^{\beta x + \gamma^* x^2} \nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta x + \gamma^* x^2} \nu_X(dx)} \tag{6.22}$$

Let this unique solution (for β) be $\beta = \beta^*$.

This is nothing, but to say conclusively that the following **equality** holds for the value β^* of β:

$$\mu_2^* = \frac{\int\limits_{\mathcal{X}_X} x^2 e^{\beta^* x + \gamma^* x^2} \nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta^* x + \gamma^* x^2} \nu_X(dx)} \tag{6.23}$$

where β^* has been uniquely determined by (6.22) saying that the following **equality** must hold

$$\mu_1^* = \frac{\int\limits_{\mathcal{X}_X} x e^{\beta^* x + \gamma^* x^2} \nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta^* x + \gamma^* x^2} \nu_X(dx)} \tag{6.24}$$

Hence, in other words, for given values of μ_1 and μ_2, namely μ_1^* and μ_2^* respectively, $\{\beta = \beta^*, \gamma = \gamma^*\}$ is the only solution of the system (6.8).

Whence, the uniqueness of the solution of (6.8) has been proved and thereby **establishing the proposition 6.3.2**. □

Remark 6.3.2. ***Importantly***, *this* ***aforesaid*** *uniqueness of the solution of (6.8) is tantamount to the uniqueness of the solution of each of the systems of simultaneous equations (11.36) and (11.57) belonging to uni-extremal cases, together with the uniqueness of the solution of (4.2) for* $m = 2$. □

6.3.4 Fourth lemma for the uni-extremity

This lemma shall establish the following proposition:

Proposition 6.3.3 ($\frac{d\beta}{d\gamma}\Big|_{\mu_1=\mu_1^*} < 0$)**.** *For any arbitrarily fixedly chosen value of* μ_1*, say* μ_1^**, such that* $0 < \mu_1^* < 1$*, the inequality* $\frac{d\beta}{d\gamma}\Big|_{\mu_1=\mu_1^*} < 0$ *holds.*

Proof of the ***proposition 6.3.3****.* By (6.11), the statement $\mu_3 - \mu_1\mu_2 > \frac{\mu_2}{\mu_1}(\mu_2 - \mu_1^*) > 0$ has already been established and therefore, by (6.15), for any fixed $\mu_1 = \mu_1^*$, the value of γ **increases monotonically** with the **monotonic decrease** in the value of β and vice versa, which is implied by

$$\frac{d\beta}{d\gamma}\Big|_{\mu_1=\mu_1^*} = -\frac{\mu_3 - \mu_1^*\mu_2}{\mu_2 - \mu_1^* 2} < 0 \tag{6.25}$$

thereby proving our **proposition 6.3.3**. □

Remark 6.3.3. ***Notably***, *this very characteristic of* $\frac{d\beta}{d\gamma}\Big|_{\mu_1=\mu_1^*} < 0$ *is used to generate the data consisting of different tables, such that each such table, which is constructed for a fixed value of* μ_1, *say* μ_1^*, *consists of records. Each such record consisting of 4 values is displayed as* $(\gamma, \beta, \mu_1^*, \mu_2)$. *These data consisting of several tables are meant for the storage of the starting values for the solution of (6.8) in the continuous case*[1]. *The generation of the said data was performed by using the program* Mathematica. *The description of the said starting values is treated in full detail subsequently in course of discussions of numerical algorithms.*

[1]As an alternative to the classification of the entire date into several tables, the entire data is alternatively stored in a long *mysql* table named as *mep*2. This *mysql* table has been made to belong to the database named as *Stochastikon*. But for our present work, we shall not use this *mysql* table, but several text-file tables instead. Each such text-file table is categorized according to the specified μ_1^*

6.3.5 Fifth lemma for the uni-extremity

For the sake of simplicity, let us take X to be a continuous random variable here (a discrete random variable X **may** cause a problem). This lemma takes care of proving the following proposition:

Proposition 6.3.4. *If μ_1^* be any fixedly chosen value of μ_1 ($0 < \mu_1^* < 1$) the **first** equation belonging to the system of equations (6.8) is given to be*

$$\mu_1^* = \frac{\int\limits_0^1 x e^{\beta x + \gamma x^2} dx}{\int\limits_0^1 e^{\beta x + \gamma x^2} dx} \tag{6.26}$$

*Then, β in this case is **uniquely determined** for any fixed value of γ and **conversely**.*

*Proof of the **proposition 6.3.4**.* By (6.9) of the **statement 6.3.1**, since we have $\frac{\partial \mu_1}{\partial \beta} > 0$, it is clear that for the fixed value of γ, μ_1 **strictly increases** monotonically with the **strict increase** of β (or equivalently, μ_1 strictly decreases monotonically with the strict decrease of β). Thus, for every chosen value of μ_1 there exists an unique value of β.

By **exactly** the **analogous argument**, by (6.11) of the **statement 6.3.3**, since we have $\frac{\partial \mu_1}{\partial \gamma} > 0$, it is clear that for any fixed value of β, μ_1 **strictly increases** monotonically with the **strict increase** of γ (or equivalently, μ_1 strictly decreases monotonically with the strict decrease of γ). Thus, for every chosen value of μ_1 there exists an unique value of γ.

Whence, for the fixed μ_1^*, (6.26) is **uniquely solvable** for β for a given fixedly chosen γ and **conversely**, (6.26) is uniquely solvable for γ for a given fixedly chosen β. This confirms our **proposition 6.3.4**. □

6.3.6 Sixth lemma for the uni-extremity

A similar lemma as the preceding one is given here. This lemma takes care of proving the following proposition:

Proposition 6.3.5. *If μ_2^* be any fixedly chosen value of μ_2 ($(\mu_1^*)^2 < \mu_2^* < \mu_1^*$) the **second** equation belonging to the system of equations (6.8) is given to be*

$$\mu_2^* = \frac{\int\limits_0^1 x^2 e^{\beta x + \gamma x^2} dx}{\int\limits_0^1 e^{\beta x + \gamma x^2} dx} \tag{6.27}$$

*Then, β in this case is **uniquely determined** for any fixed value of γ and **conversely**.*

*Proof of the **proposition 6.3.5**.* **If** γ has to be computed for a fixedly chosen β, then by (6.10) of the **statement 6.3.2**, since we have $\frac{\partial \mu_2}{\partial \gamma} > 0$, for every chosen value of μ_2 there exists an unique value of γ.

If β has to be computed for a fixedly chosen γ, then by (6.11) of the **statement 6.3.3**, since we have $\frac{\partial \mu_2}{\partial \beta} > 0$, for every chosen value of μ_2 there exists an unique value of β.

Whence, for the fixed μ_2^*, (6.27) is **uniquely solvable** for β for a given fixedly chosen γ and conversely, (6.27) is uniquely solvable for γ for a given fixedly chosen β. This confirms our **proposition 6.3.5**. □

6.3.7 Seventh lemma for the uni-extremity

In this lemma, we shall prove that, with subject to the consideration of the system of simultaneous equations (6.34) (i.e. the continuous case of X for the system (6.8)), the **necessary** and the **sufficient** condition for $\mu_1 = \frac{1}{2}$ is $\beta + \gamma = 0$.

In other words, we need to prove the following proposition:

Proposition 6.3.6 (The necessary and sufficient condition for $\mu_1 = \frac{1}{2}$). *In the **light** of the system of simultaneous equations (6.34),*

$$\beta + \gamma = 0 \Longleftrightarrow \mu_1 = \frac{1}{2} \tag{6.28}$$

Proof of the ***proposition 6.3.6***. For to show that $\mu_1 = \frac{1}{2} \Longleftarrow \beta + \gamma = 0$, we have

$$\begin{aligned} E[X] &= \frac{\int_0^1 x e^{\beta x - \beta x^2} dx}{\int_0^1 e^{\beta x - \beta x^2} dx} \qquad (\text{ by setting } \gamma = -\beta) \\ &= \frac{\left(e^{\beta x} \int x e^{-\beta x^2} dx - \beta \int e^{\beta x} \frac{1}{-2\beta} e^{-\beta x^2} dx \right) \Big|_{x=0}^{x=1}}{\int_0^1 e^{\beta x - \beta x^2} dx} \qquad (6.29) \\ &= \frac{\left(e^{\beta x} \left(\frac{1}{-2\beta} e^{-\beta x^2} \right) \right) \Big|_{x=0}^{x=1} + \frac{1}{2} \int_0^1 e^{\beta x - \beta x^2} dx}{\int_0^1 e^{\beta x - \beta x^2} dx} \\ &= \frac{1}{2} = \mu_1 \end{aligned}$$

and for to show that, $\mu_1 = \frac{1}{2} \Longrightarrow \beta + \gamma = 0$ the argument is rather simple:

By the **proposition 6.3.4**, for any fixed value of μ_1, say $\mu_1 = \mu_1^*$, the equation (6.26) is **uniquely solvable** for β corresponding to any **fixedly predetermined** value of γ.

Accordingly, corresponding to $\mu_1^* = \frac{1}{2}$ and by choosing $\gamma = \gamma_0$, β must be **uniquely determinable**.

Hence, since $\beta = -\gamma_0$ already satisfies (6.29), $\beta = -\gamma_0$ is the **unique** solution of (6.26) with subject to the given $\gamma = \gamma_0$ corresponding to $\mu_1^* = \frac{1}{2}$.

Thus, $\mu_1 = \frac{1}{2} \Longrightarrow \beta + \gamma = 0$.

Whence, $\mu_1 = \frac{1}{2} \Longleftrightarrow \beta + \gamma = 0$ and our **proposition 6.3.6** thereby gets proved. □

6.3.8 Eighth lemma for the uni-extremity

This lemma shall show that for the continuous case of X, the following proposition holds:

Proposition 6.3.7 (Condition for the symmetry). *The* ***necessary*** *and* ***sufficient*** *condition for the probability density curve for X to be* ***symmetric*** *(about the middle point of the support, namely $x = \frac{1}{2}$) is $\mu_1 = \frac{1}{2}$ (or* ***equivalently*** *$\beta + \gamma = 0$)*

Proof of the ***proposition 6.3.7****.* This assertion can be proved by taking the probability density function of X to be $f_{X|\{(\frac{1}{2},\mu_2)\}}(x) = Ke^{\beta x+\gamma x^2}$, $0 \leq x \leq 1$.

By the symmetric property of the probability density curve $f_{X|\{(\frac{1}{2},\mu_2)\}}(x)$ about the middle point $x = \frac{1}{2}$, we get

$$
\begin{aligned}
& f_{X|\{(\frac{1}{2},\mu_2)\}}\left(\frac{1}{2}+\xi\right) = f_{X|\{(\frac{1}{2},\mu_2)\}}\left(\frac{1}{2}-\xi\right), \quad \xi \in \left[0, \frac{1}{2}\right] \\
& \Leftrightarrow \beta\left(\frac{1}{2}-\xi\right) + \gamma\left(\frac{1}{2}-\xi\right)^2 = \beta\left(\frac{1}{2}+\xi\right) + \gamma\left(\frac{1}{2}+\xi\right)^2 \\
& \Leftrightarrow \beta(-\xi-\xi) = \gamma(\xi+\xi) \\
& \Leftrightarrow \beta = -\gamma
\end{aligned}
$$

and by the very assertion $\mu_1 = \frac{1}{2} \Longleftrightarrow \beta + \gamma = 0$ of the **proposition 6.3.6**, we establish that the necessary and sufficient condition for the symmetry of the probability density curve of X is $\mu_1 = \frac{1}{2}$ or equivalently $\beta + \gamma = 0$.

This completes the proof of our **proposition 6.3.7**. □

6.3.9 Ninth lemma for the uni-extremity

As a special case for $m = 2$, this lemma **reestablishes** the uniqueness of the solution of the system (6.8) is tantamount to the positive definiteness of the following symmetric matrix

$$\begin{pmatrix} \frac{\partial \mu_1}{\partial \beta} & \frac{\partial \mu_1}{\partial \gamma} \\ \frac{\partial \mu_2}{\partial \beta} & \frac{\partial \mu_2}{\partial \gamma} \end{pmatrix} \tag{6.30}$$

The symmetry of this positive definite matrix, with subject to $0 < \mu_1 < 1$ and $\mu_1^2 < \mu_2 < \mu_1$, is evident by the very facts that are presented as follows:

- With the help of (6.9), we get $\frac{\partial \mu_1}{\partial \beta} = \mu_2 - \mu_1^2 > 0$, simply because of

$$\mu_2 - \mu_1^2 = \sigma^2 > 0 \tag{6.31}$$

- With the help of (6.9), (6.10) and (6.11) on the established inequality, namely

$$\begin{aligned} &(\mu_4 - \mu_2^2)(\mu_2 - \mu_1^2) - (\mu_3 - \mu_1\mu_2)^2 \\ =&(\mu_4 - \mu_2^2)(\mu_2 - \mu_1^2)\left(1 - \frac{(\mu_3 - \mu_1\mu_2)^2}{(\mu_4 - \mu_2^2)(\mu_2 - \mu_1^2)}\right) \\ =&\sigma_{X^2}^2\sigma^2\left(1 - \rho_{X,X^2}^2\right) > 0 \end{aligned} \tag{6.32}$$

we get $\begin{vmatrix} \frac{\partial \mu_1}{\partial \beta} & \frac{\partial \mu_1}{\partial \gamma} \\ \frac{\partial \mu_2}{\partial \beta} & \frac{\partial \mu_2}{\partial \gamma} \end{vmatrix} = (\mu_4 - \mu_2^2)(\mu_2 - \mu_1^2) - (\mu_3 - \mu_1\mu_2)^2 > 0$ and

- With the help of (6.11), $\frac{\partial \mu_1}{\partial \gamma} = \frac{\partial \mu_2}{\partial \beta} = \mu_3 - \mu_1\mu_2 > 0$

Owing to the very fact that the inequalities (6.31) and (6.32) do fulfill the basically needed requirement for the proof of the uniqueness of the solution of the system (6.8), we arrive at the following conclusion

> the positive definiteness of the matrix (6.30) is simply an equivalent statement to the uniqueness of the solution of the system (6.8). (6.33)

6.3.10 The existence of the solution of equation-system for $m = 2$

Theorem 6.3.1 (Existence for $m = 2$). *Even in this case for $m = 2$, for the sake of convenience, let us take X for a* ***continuous*** *random variable at first. In that case, we get (6.8) as*

$$\begin{cases} \mu_1 &= \dfrac{\int\limits_0^1 x e^{\beta x+\gamma x^2} dx}{\int\limits_0^1 e^{\beta x+\gamma x^2} dx} \\ \\ \mu_2 &= \dfrac{\int\limits_0^1 x^2 e^{\beta x+\gamma x^2} dx}{\int\limits_0^1 e^{\beta x+\gamma x^2} dx} \end{cases} \tag{6.34}$$

Then, the solution of (6.34) ***exists*** *and the* ***same*** *is* ***not*** *fully the case for a* ***discrete*** *X.*

Proof of the ***theorem 6.3.1****.* Basically, our objective is to show that the images corresponding to all the possible real values of β and γ (i.e. $\beta, \gamma \in \mathbb{R}$) span the entire region

$$R_{\mu_1,\mu_2} = \{(\mu_1, \mu_2) \mid 0 < \mu_1 < 1, \mu_1^2 < \mu_2 < \mu_1\}$$

where the geometrical figure of the region R_{μ_1,μ_2} is **given in the section 5.10**. This figure R_{μ_1,μ_2} has also been referred to the page 74 of [25].

The positivity of the partial derivatives (6.9), (6.10) and (6.11) belonging to the **statements 6.3.1**, **6.3.2** and **6.3.3** (of **lemma 1**) respectively, namely

$$\frac{\partial \mu_1}{\partial \beta} = \mu_2 - \mu_1^2 = \sigma^2 > 0 \tag{6.35}$$

$$\frac{\partial \mu_2}{\partial \gamma} = \mu_4 - \mu_2^2 = \sigma_{X^2}^2 > 0 \tag{6.36}$$

$$\frac{\partial \mu_1}{\partial \gamma} = \frac{\partial \mu_2}{\partial \beta} = \mu_3 - \mu_1\mu_2 = Cov[X^2, X] > 0 \tag{6.37}$$

are basically explained by the **covariances** between X^i and X^j for $i, j \in \{1, 2\}$. In other words, the covariances are bound to be **positive**.

These **restatements** (6.35), (6.36) and (6.37) shall play an important in our subsequent steps.

The very assertion with regard to the fact that the images corresponding to the values $(\beta, \gamma) \in \mathbb{R}^2$ span the entire R_{μ_1,μ_2} shall be proved in a few steps:

Step 1:

In this step, we shall show the following:

1. the change of μ_2 is strict monotonic both with the changes of β and γ individually.
2. as μ_2 approaches its **greatest lower bound** μ_1^{*2}, then $\gamma \to -\infty$ and $\beta \to +\infty$ and vice versa, i.e.

$$\mu_2 \to \mu_1^{*2} \Longleftrightarrow (\gamma \to -\infty \,\&\, \beta \to +\infty) \tag{6.38}$$

   In other words, the **parabolic arc** OR_1QP of the geometrical figure of R_{μ_1,μ_2} described by the equation $\mu_2 = \mu_1^2$, **with the exception of the points** O **and** P, represents the points $(\beta = +\infty, \gamma = -\infty)$.

3. as μ_2 approaches its **least upper bound** μ_1^*, then $\gamma \to +\infty$ and $\beta \to -\infty$ and vice versa, i.e.

$$\mu_2 \to \mu_1^* \Longleftrightarrow (\gamma \to +\infty \,\&\, \beta \to -\infty) \tag{6.39}$$

   In other words, the **line segment** OR_2RP of the geometrical figure of R_{μ_1,μ_2} described by the equation $\mu_2 = \mu_1$, **with the exception of the points** O **and** P, represents the points $(\beta = -\infty, \gamma = +\infty)$.

Now, by (6.25), for any fixed value μ_1^* of μ_1, we already have

$$\left.\frac{d\beta}{d\gamma}\right|_{\mu_1=\mu_1^*} = -\frac{\mu_3 - \mu_1^*\mu_2}{\mu_2 - \mu_1^*2} < 0 \text{ (referred to the previous } \textbf{(6.25)}\text{)}$$

which clearly shows that the increase of β is **strict monotonic** with the **strict monotonic** decrease in γ for any fixed μ_1^*.

Again, by (6.16) for the fixed μ_1^*, namely

$$\left.\frac{d\mu_2}{d\gamma}\right|_{\mu_1=\mu_1^*} = \sigma_{X^2}^2\left(1 - \rho_{X,X^2}^2\right) > 0 \text{ (referred to the previous } \textbf{(6.16)}\text{)}$$

we get by combining (6.25) and (6.16) for the fixed μ_1^* as

$$\left.\frac{d\mu_2}{d\beta}\right|_{\mu_1=\mu_1^*} = \frac{\left.\frac{d\mu_2}{d\gamma}\right|_{\mu_1=\mu_1^*}}{\left.\frac{d\beta}{d\gamma}\right|_{\mu_1=\mu_1^*}} < 0 \tag{6.40}$$

and thus, for the fixed μ_1^*, we have the following

- by (6.16), the **increase** of μ_2 is **strict monotonic** with the **strict monotonic** increase in γ and vice versa.
- by (6.40), the **decrease** of μ_2 is **strict monotonic** with the **strict monotonic** increase in β and vice versa.

thereby proving our first assertion of this step.

Now, we shall analyze the two cases $\mu_2 \to \mu_1^{*2}$ and $\mu_2 \to \mu_1^*$ **geometrically**. Additionally, with reference to the **Definition and Proposition 4.4.2**, the probability distributions subjecting to the two cases of $\mu_2 = \mu_1^{*2}$ and $\mu_2 = \mu_1^*$ are pure and simple **discrete** probability distributions. As a matter of fact, these discrete probability distributions are well conventionally known **degenerated** and **Bernoulli** probability distributions.

The Case of $\mu_2 \to \mu_1^{*2}$ (from right):

By referring to the knowledge of the standard normal probability distribution that has been discussed in (11.28), we have

$$\frac{1}{\sqrt{2\pi}}\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1^*}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}}\int_{-\frac{\mu_1^*}{\sigma}}^{\frac{1-\mu_1^*}{\sigma}} e^{-\frac{x^2}{2}} dx \overset{\sigma\to 0}{\longrightarrow} 1$$

and therefore with subject to

$$1 = \int_0^1 e^{\alpha+\beta x+\gamma x^2} dx \approx \frac{1}{\sqrt{2\pi}}\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1^*}{\sigma}\right)^2} dx \tag{6.41}$$

by comparing the coefficients of x and x^2 situated in the power of e of both sides of the above relation (6.41), we get $\beta \approx \frac{\mu_1^*}{\sigma^2}$ and $\gamma \approx -\frac{1}{2\sigma^2}$.

Notably, the values of $\int_0^1 e^{\alpha+\beta x+\gamma x^2}dx = 1$ and $\frac{1}{\sqrt{2\pi}}\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1^*}{\sigma}\right)^2}dx$ stated in (6.41) get **infinitesimally** closer to each other, if the value of σ is made to get **infinitesimally** closer to 0.

This is to say that β and γ get **infinitesimally** closer to the values $\frac{\mu_1^*}{\sigma^2}$ and $-\frac{1}{2\sigma^2}$ respectively, if the value of $\sigma = \sqrt{\mu_2 - \mu_1^{*2}}$ is decreased **arbitrarily**.

This proves nothing, but the very fact that $\mu_2 \to \mu_1^{*2}$ implies and implied by $\beta \to +\infty$ and $\gamma \to -\infty$, thereby proving our second assertion of this step. Notably, the behavior, in which $\beta \to +\infty$ and $\gamma \to -\infty$, strictly depends on the value of μ_1^*.

Geometrically, this aforesaid behavior is described by the fact that the peakedness of the bell shaped probability density curve gets arbitrarily higher (i.e. the ordinate of the probability density function at the point $x = \mu_1^*$ gets arbitrarily larger). In the limiting sense for $\mu_2 \to \mu_1^{*2}$, the **probability distribution** is **discrete**, the probability mass function of which being given by $f_{X|\{(\mu_1^*,\mu_1^{*2})\}}(x) = 1$, $x = \mu_1^*$ and thereby affirming the **Definition and Proposition 4.4.2**.

From the analytical angle, for the fixed μ_1^*, under the consideration of the following:

- corresponding to each input (μ_1^*, μ_2), (β, γ) can only be uniquely determined as a solution of (6.34) (referred to the **proposition 6.3.2** (of **lemma 3**)).
- the **strict monotone decrease** of μ_2 for $\mu_2 \to \mu_1^{*2}+$ with
  - **strict monotone decrease** of γ **confirmed by (6.20)**, i.e. by $\frac{d\mu_2}{d\gamma}\Big|_{\mu_1=\mu_1^*} > 0$
  - **strict monotone increase** of β **confirmed by (6.40)**, i.e. by $\frac{d\mu_2}{d\beta}\Big|_{\mu_1=\mu_1^*} < 0$

for any given **arbitrarily** small ϵ ($\epsilon > 0$) there exist two positive numbers G_β and G_γ, such that $\left|\mu_2 - \mu_1^{*2}\right| < \epsilon$ for every $\beta > G_\beta$ and $\gamma < -G_\gamma$, but with subject to the **fulfillment** of (6.26).

Now, since by the **proposition 6.3.4** (of **lemma 5**), for the fixed $\mu_1 = \mu_1^*$, β is uniquely determinable by γ (or conversely) and by the **proposition 6.3.3** (of **lemma 4**), namely $\frac{d\beta}{d\gamma}\Big|_{\mu_1=\mu_1^*} < 0$, it is evidently clear that the arbitrariness of the largeness of β ($\beta > 0$) **implies and implied by** the arbitrariness of the largeness of $-\gamma$ ($-\gamma > 0$).

In other words, the closeness of μ_2 to μ_1^{*2} **implies and implied by** the arbitrariness of the largeness of the positivity of β (or **equivalently** of $-\gamma$). This means nothing, but $\mu_2 \to \mu_1^{*2} \iff (\gamma \to -\infty \,\&\, \beta \to +\infty)$ and thereby establishing the assertion (6.38).

The Case of $\mu_2 \to \mu_1^*$ (from left):

We have already proved that if μ_2 is made to increase monotonically, then β monotonically decreases and γ monotonically increases.

So, the question arises, what could happen to the **decrease of** β and the **increase of** γ, if μ_2 is made to increase to bring **arbitrarily close to** μ_1^*.

As a matter of fact, if μ_2 made to get close to μ_1^* from left beyond a limit, the probability density function of X becomes a bathtub shaped (*this point shall be discussed in full details in due course*). **Geometrically**, this bathtub shaped probability density curve of X becomes **flatter and flatter and tend to take the shape of a rectangular geometrical trough with two bottom-edges** (i. e. **arbitrarily flatter**) with the **infinitesimal closeness** of μ_2 to μ_1^* from left and this happens, purely when $\beta \to -\infty$ and $\gamma \to +\infty$.

This proves nothing, but the very fact that $\mu_2 \to \mu_1^*$ implies and implied by $\beta \to -\infty$ and $\gamma \to +\infty$, thereby proving our third assertion of this step. Even in this case, notably, the behavior, in which $\beta \to -\infty$ and $\gamma \to +\infty$, strictly depends on the value of μ_1^*. In the limiting sense for $\mu_2 \to \mu_1^*$, the **probability distribution** is **discrete**, the probability mass function of which being given by $f_{X|\{(\mu_1^*,\mu_1^*)\}}(x) = \begin{cases} 1-\mu_1^* & : \quad x = 0 \\ \mu_1^* & : \quad x = 1 \end{cases}$ and thereby affirming the **Definition and Proposition 4.4.2**.

From the analytical angle, for the fixed μ_1^*, under the consideration of

the following:

- corresponding to each input (μ_1^*, μ_2), (β, γ) can only be uniquely determined as a solution of (6.34) (referred to the **proposition 6.3.2** (of the **lemma 3**)).
- the **strict monotone increase** of μ_2 for $\mu_2 \to \mu_1^*-$ with
  - **strict monotone increase** of γ confirmed by (6.20), i.e. by $\frac{d\mu_2}{d\gamma}\Big|_{\mu_1=\mu_1^*} > 0$
  - **strict monotone decrease** of β confirmed by (6.40), i.e. by $\frac{d\mu_2}{d\beta}\Big|_{\mu_1=\mu_1^*} < 0$

for any given **arbitrarily** small ϵ ($\epsilon > 0$) there exist two positive numbers G_β and G_γ, such that $|\mu_2 - \mu_1^*| < \epsilon$ for every $\beta < -G_\beta$ and $\gamma > G_\gamma$, but with subject to the **fulfillment** of (6.26).

Now, since by the **proposition 6.3.4** (of **lemma 5**), for the fixed $\mu_1 = \mu_1^*$, γ is uniquely determinable by β (or conversely) and by the **proposition 6.3.3** (of **lemma 4**), namely $\frac{d\beta}{d\gamma}\Big|_{\mu_1=\mu_1^*} < 0$, it is evidently clear that the arbitrariness of the largeness of γ ($\gamma > 0$) **implies and implied by** the arbitrariness of the largeness of $-\beta$ ($-\beta > 0$).

In other words, the closeness of μ_2 to μ_1^* **implies and implied by** the arbitrariness of the largeness of the positivity of γ (or **equivalently** of $-\beta$). This means nothing, but $\mu_2 \to \mu_1^* \iff (\gamma \to +\infty \,\&\, \beta \to -\infty)$ and thereby establishing the assertion (6.39).

The conclusive statement of this **first** step:

Conclusively, if μ_2 is made to increase strictly monotonically from μ_1^{*2} to μ_1^*, then

- β **decreases strictly monotonically** from $+\infty$ to $-\infty$
- γ **increases strictly monotonically** from $-\infty$ to $+\infty$

Step 2:

Elementarily, with subject to the defined relationship (6.34) between (μ_1, μ_2) and (β, γ), the evaluated values of (μ_1, μ_2) for every given $(\beta, \gamma) \in \mathbb{R}^2$ **must be contained in** R_{μ_1,μ_2}.

In this step, we shall show that, for all the images (μ_1, μ_2) of (β, γ) defined by (6.34) falling within the bounded region R_{μ_1,μ_2}, for the fixedly chosen μ_1^*,

1. $\beta + \gamma < 0$ for $\mu_1^* < \frac{1}{2}$
2. $\beta + \gamma = 0$ for $\mu_1^* = \frac{1}{2}$
3. $\beta + \gamma > 0$ for $\mu_1^* > \frac{1}{2}$

Basically, with the help of the very **established proposition 6.3.6 (of the lemma 7)**, namely

$$\mu_1^* = \frac{1}{2} \iff \beta + \gamma = 0 \tag{6.42}$$

we shall go ahead to prove that $\mu_1^* \lesseqgtr \frac{1}{2}$ according as $\beta + \gamma \lesseqgtr 0$. This can be proved in two ways and we shall give both these proofs for the sake of **completeness** and **complete transparency**. Of course, each of these proofs is small enough to be presented beautifully and transparently.

For this, we shall set $\beta + \gamma = \epsilon$, ϵ being any real value. With this, we proceed as follows:

- By (6.35), namely $\frac{\partial \mu_1}{\partial \beta} > 0$, μ_1 is **strictly monotonic increasing** (or **decreasing**) with respect to β for any fixedly chosen value of γ, say for $\gamma = \gamma_0$.

  therefore, by $\beta + \gamma_0 = \epsilon$, i. e. by putting $\beta = -\gamma_0 + \epsilon$, we arrive at

$$\mu_1 = \frac{\int_0^1 x e^{(-\gamma_0+\epsilon)x+\gamma_0 x^2} dx}{\int_0^1 e^{(-\gamma_0+\epsilon)x+\gamma_0 x^2} dx} \quad > \quad \frac{1}{2}, \text{ if } \epsilon > 0, \text{ by monotonic } \textbf{increasing}$$

$$< \quad \frac{1}{2}, \text{ if } \epsilon < 0, \text{ by monotonic } \textbf{decreasing}$$

$$= \quad \frac{1}{2}, \text{ if } \epsilon = 0, \text{ by (6.42)}$$

- **Alternatively**, by (6.37), namely $\frac{\partial \mu_1}{\partial \gamma} > 0$, μ_1 is **strictly monotonic increasing** (or **decreasing**) with respect to γ for any fixedly chosen value of β, say for $\beta = \beta_0$.

  therefore, by $\beta_0 + \gamma = \epsilon$, i. e. by putting $\gamma = -\beta_0 + \epsilon$, we arrive at

$$\mu_1 = \frac{\int_0^1 x e^{\beta_0 x+(-\beta_0+\epsilon)x^2} dx}{\int_0^1 e^{\beta_0 x+(-\beta_0+\epsilon)x^2} dx} \quad > \quad \frac{1}{2}, \text{ if } \epsilon > 0, \text{ by monotonic } \textbf{increasing}$$

$$< \quad \frac{1}{2}, \text{ if } \epsilon < 0, \text{ by monotonic } \textbf{decreasing}$$

$$= \quad \frac{1}{2}, \text{ if } \epsilon = 0, \text{ by (6.42)}$$

This makes evidently clear that $\beta + \gamma > 0$ for $\mu_1^* > \frac{1}{2}$ and $\beta + \gamma < 0$ for $\mu_1^* < \frac{1}{2}$, thereby proving our assertion of this step, which formally reads

$$\mu_1^* \lessgtr \frac{1}{2} \text{ according as } \beta + \gamma \lessgtr 0 \tag{6.43}$$

The conclusive statement of this **second** step:

We have therefore proved that **all the possible signs** of the expression $\beta+\gamma$ (i.e. $\lessgtr 0$) are included for our present consideration. This is to say that the clear relationship between $\beta + \gamma$ and μ_1^* is thereby established.

Step 3:

The **two** important two-dimensional curves denoted by $C_{\beta=0}$ and $C_{\gamma=0}$, which **lie fully within the region** R_{μ_1,μ_2}, play decisive roles.

Our objective in this step shall be to examine, how do these curves $C_{\beta=0}$ and $C_{\gamma=0}$ behave with **different values** of μ_1^*, thereby **the geometrical side of the behaviors of both** $C_{\beta=0}$ **and** $C_{\gamma=0}$ **lying within** $\overline{R}_{\mu_1,\mu_2}$ **can be well described analytically**.

In view of the **geometrical figure** giving the region R_{μ_1,μ_2} given in the last page of the previous chapter, let us define the curves $C_{\beta=0}$ and $C_{\gamma=0}$ as follows:

Definition 6.3.1 (The curve $C_{\beta=0}$). *The curve $C_{\beta=0}$, which is a **dashed curve**, is denoted by $OB_2R_0B_1P$.*

With regard to the equation-system (6.34), the curve $C_{\beta=0}$ represents the set of (μ_1, μ_2) values corresponding to $\beta = 0, \gamma \in \mathbb{R}$, whose parametric representation is given as

$$\begin{cases} \mu_1 = \dfrac{\int\limits_0^1 xe^{\gamma x^2}dx}{\int\limits_0^1 e^{\gamma x^2}dx} \\ \mu_2 = \dfrac{\int\limits_0^1 x^2e^{\gamma x^2}dx}{\int\limits_0^1 e^{\gamma x^2}dx} \end{cases} \tag{6.44}$$

Definition 6.3.2 (The curve $C_{\gamma=0}$). *The curve $C_{\gamma=0}$, which is a **lined curve**, is denoted by $OC_2R_0C_1P$.*

With regard to the equation-system (6.34), the curve $C_{\gamma=0}$ represents the set of (μ_1, μ_2) values corresponding to $\beta \in \mathbb{R}, \gamma = 0$, whose parametric representation is given as

$$\begin{cases} \mu_1 = \dfrac{\int\limits_0^1 xe^{\beta x}dx}{\int\limits_0^1 e^{\beta x}dx} \\ \mu_2 = \dfrac{\int\limits_0^1 x^2e^{\beta x}dx}{\int\limits_0^1 e^{\beta x}dx} \end{cases} \tag{6.45}$$

These **two curves** $C_{\beta=0}$, $C_{\gamma=0}$ and the **line segment** $\overline{R_2R_0R_1}$ divide the **closed region** $\overline{R}_{\mu_1,\mu_2}$ (the region $\overline{R}_{\mu_1,\mu_2}$ includes the **boundary points** of R_{μ_1,μ_2}) broadly into **eight** subdivisions.

These subdivisions are termed as **subregions**, in each of which the signs of β, γ and $\beta+\gamma$ are individually specified. **The detailed description of these subregions are given individually in the step 4**.

For the study of the **geometrical behaviors** of $C_{\beta=0}$ and $C_{\gamma=0}$ (*the behaviors are already presented in the geometrical figure of* R_{μ_1,μ_2}), we need to distinguish the **three cases**, namely

- $\mu_1^* < \frac{1}{2}$
- $\mu_1^* > \frac{1}{2}$ and
- $\mu_1^* = \frac{1}{2}$

Before we proceed to discuss **these three cases** one by one, let us denote β_0 and γ_0 in the following way:

- β_0 denotes the unique solution of $\mu_1^* = \frac{\int_0^1 xe^{\beta x}dx}{\int_0^1 e^{\beta x}dx}$.

  Note, for the fixed $\gamma = 0$, by (6.35), β_0 is unique.

  Furthermore, let $\mu_2^{(\gamma=0)} = \frac{\int_0^1 x^2e^{\beta_0 x}dx}{\int_0^1 e^{\beta_0 x}dx}$

- γ_0 denotes the unique solution of $\mu_1^* = \frac{\int_0^1 xe^{\gamma x^2}dx}{\int_0^1 e^{\gamma x^2}dx}$.

  Note, for the fixed $\beta = 0$, by (6.37), γ_0 is unique.

  Furthermore, let $\mu_2^{(\beta=0)} = \frac{\int_0^1 x^2e^{\gamma_0 x^2}dx}{\int_0^1 e^{\gamma_0 x^2}dx}$

The case of $\mu_1^* < \frac{1}{2}$:

By considering the curve $C_{\gamma=0}$, by (6.7), we get $(\beta_0 < 0, \gamma = 0)$ at the point $B_2 : (\mu_1^*, \mu_2^{(\gamma=0)})$.

Again, by considering the curve $C_{\beta=0}$ at the point $C_2 : (\mu_1^*, \mu_2^{(\beta=0)})$, the corresponding value of (β, γ) is $(\beta = 0, \gamma_0)$. The question is, what should be the sign of γ_0?

We see very clearly, that for the fixed μ_1^*, if μ_2 is made to change from $\mu_2^{(\gamma=0)}$ to $\mu_2^{(\beta=0)}$ (namely from the position B_2 to the position C_2), then β would change itself from $\beta_0(< 0)$ to 0.

This means, β has **increased** from $\beta_0(< 0)$ to 0, while μ_2 has changed itself from $\mu_2^{(\gamma=0)}$ to $\mu_2^{(\beta=0)}$.

Therefore,

- by (6.40), μ_2 has **decreased** from $\mu_2^{(\gamma=0)}$ to $\mu_2^{(\beta=0)}$, i.e. $\mu_2^{(\gamma=0)} > \mu_2^{(\beta=0)}$
- and by (6.25), γ has **decreased** from 0 to γ_0, i.e. $\gamma_0 < 0$

Thus, it is evidently and **conclusively** clear that, within the range of $\mu_1^* < \frac{1}{2}$, **the curve $C_{\gamma=0}$ lies above the curve $C_{\beta=0}$.**

The case of $\mu_1^* > \frac{1}{2}$:

By considering the curve $C_{\gamma=0}$, by (6.7), we get $(\beta_0 > 0, \gamma = 0)$ at the point $B_1 : (\mu_1^*, \mu_2^{(\gamma=0)})$.

Again, by considering the curve $C_{\beta=0}$ at the point $C_1 : (\mu_1^*, \mu_2^{(\beta=0)})$, the corresponding value of (β, γ) is $(\beta = 0, \gamma_0)$. The question is, what should be the sign of γ_0?

We see very clearly, that for the fixed μ_1^*, if μ_2 is made to change from $\mu_2^{(\gamma=0)}$ to $\mu_2^{(\beta=0)}$ (namely from the position B_1 to the position C_1), then β would change itself from $\beta_0(> 0)$ to 0.

This means, β has **decreased** from $\beta_0(> 0)$ to 0, while μ_2 has changed itself from $\mu_2^{(\gamma=0)}$ to $\mu_2^{(\beta=0)}$.

Therefore,

- by (6.40), μ_2 has **increased** from $\mu_2^{(\gamma=0)}$ to $\mu_2^{(\beta=0)}$, i.e. $\mu_2^{(\gamma=0)} < \mu_2^{(\beta=0)}$
- and by (6.25), γ has **increased** from 0 to γ_0, i.e. $\gamma_0 > 0$

Thus, it is evidently and **conclusively** clear that, within the range of $\mu_1^* > \frac{1}{2}$, **the curve $C_{\gamma=0}$ lies below the curve $C_{\beta=0}$.**

The case of $\mu_1^* = \frac{1}{2}$:

By considering the curve $C_{\gamma=0}$, by (6.7), we get $(\beta_0 = 0, \gamma = 0)$ at the point $R_0 : (\frac{1}{2}, \mu_2^{(\gamma=0)})$.

Trivially, in this case, $\mu_2^{(\gamma=0)} = \mu_2^{(\beta=0)} = \frac{1}{3}$ and thus the curves $C_{\gamma=0}$ and $C_{\beta=0}$ **intersect each other** at the point $R_0 : (\frac{1}{2}, \frac{1}{3})$.

The summarized conclusion of this **third** step:

- Within the range of $\mu_1^* < \frac{1}{2}$, the curve $C_{\gamma=0}$ lies **above** the curve $C_{\beta=0}$
- Within the range of $\mu_1^* > \frac{1}{2}$, the curve $C_{\gamma=0}$ lies **below** the curve $C_{\beta=0}$
- At $\mu_1^* = \frac{1}{2}$, the curves $C_{\gamma=0}$ **intersects** the curve $C_{\beta=0}$ (at the point $R_0 : (\frac{1}{2}, \frac{1}{3})$)

This completes the discussions of the curves $C_{\gamma=0}$ and $C_{\beta=0}$ with subject to the cases $\mu_1^* \lessgtr \frac{1}{2}$.

Step 4:

In this step, we shall discuss about the signs of β, γ and $\beta + \gamma$ in different (bounded) subregions of $\overline{R}_{\mu_1,\mu_2}$ **with the exception of the corner points** O **and** P. For this,

- we shall use the monotonic character of μ_2 with respect to γ and β individually for the fixed μ_1^*, namely the (6.20) saying $\frac{d\mu_2}{d\gamma} > 0$ and the (6.40) saying $\frac{d\mu_2}{d\beta} < 0$
- we shall use the deduced statements in the step 1, namely
  - the (6.38) saying $\mu_2 \to \mu_1^{*2} \Longleftrightarrow (\gamma \to -\infty \,\&\, \beta \to +\infty)$
  - the (6.39) saying $\mu_2 \to \mu_1^* \Longleftrightarrow (\gamma \to +\infty \,\&\, \beta \to -\infty)$

Moreover, it has to be kept in mind that

- any point of the curve $C_{\gamma=0}$ is the **point of change in sign** of γ.
- any point of the curve $C_{\beta=0}$ is the **point of change in sign** of β.

With this, we proceed to discuss the different subregions one by one:

1. The subregion for $\beta \leq \mathbf{0}, \gamma \leq \mathbf{0}, \beta + \gamma \leq \mathbf{0}$:

This subregion $OB_2R_0C_2O$ is bounded by the curves $C_{\gamma=0}$ and $C_{\beta=0}$.

Since the curve $C_{\gamma=0}$ lies **above** the curve $C_{\beta=0}$ in this case, by the monotonic character of μ_2, namely (6.20) and (6.40),

- any point lying **below** $C_{\gamma=0}$ must correspond to $\gamma < 0$
- any point lying **above** $C_{\beta=0}$ must correspond to $\beta < 0$

This shows that the **interior** of this subregion consists of points corresponding to $\beta < 0, \gamma < 0$ and $\beta + \gamma < 0$.

2. The subregion for $\beta \geq \mathbf{0}, \gamma \geq \mathbf{0}, \beta + \gamma \geq \mathbf{0}$:

This subregion $PB_1R_0C_1P$ is bounded by the curves $C_{\gamma=0}$ and $C_{\beta=0}$.

Since the curve $C_{\gamma=0}$ lies **below** the curve $C_{\beta=0}$ in this case, by the monotonic character of μ_2, namely (6.20) and (6.40),

- any point lying **above** $C_{\gamma=0}$ must correspond to $\gamma > 0$
- any point lying **below** $C_{\beta=0}$ must correspond to $\beta > 0$

This shows that the **interior** of this subregion consists of points corresponding to $\beta > 0, \gamma > 0$ and $\beta + \gamma > 0$.

3. The subregion for $\beta \leq 0, \gamma \geq 0, \beta + \gamma \leq 0$:

This subregion $OR_2R_0B_2O$ is bounded by the line segments $\overline{OR_2}$ and $\overline{R_2R_0}$ and the curve $C_{\gamma=0}$.

In this case, by the monotonic character of μ_2, namely (6.20) and (6.40),

- any point lying **above** $C_{\gamma=0}$ must correspond to $\gamma > 0$
- any point lying **above** $C_{\beta=0}$ must correspond to $\beta < 0$

This shows that the **interior** of this subregion consists of points corresponding to $\beta < 0, \gamma > 0$ and $\beta + \gamma < 0$.

Moreover, by the statement (6.39), the line segment $\overline{OR_2}$ of this subregion, **with the exception of the point** O, consists of points corresponding to $\beta = -\infty, \gamma = +\infty$, but $\beta + \gamma \leq 0$.

Conclusively, the points of this subregion, **with the exception of the point** O, basically correspond to $-\infty \leq \beta \leq 0, 0 \leq \gamma \leq +\infty$, but $\beta + \gamma \leq 0$.

4. The subregion for $\beta \geq 0, \gamma \leq 0, \beta + \gamma \leq 0$:

This subregion $OR_1R_0C_2O$ is bounded by the line segment $\overline{R_1R_0}$, the curve $C_{\beta=0}$ and the arc OR_1.

In this case, by the monotonic character of μ_2, namely (6.20) and (6.40),

- any point lying **below** $C_{\gamma=0}$ must correspond to $\gamma < 0$
- any point lying **below** $C_{\beta=0}$ must correspond to $\beta > 0$

This shows that the **interior** of this subregion consists of points corresponding to $\beta > 0, \gamma < 0$ and $\beta + \gamma < 0$.

Moreover, by the statement (6.38), the curve OR_1 of this subregion, **with the exception of the point** O, consists of points corresponding to $\beta = +\infty, \gamma = -\infty$, but $\beta + \gamma \leq 0$.

Conclusively, the points of this subregion, **with the exception of the point** O, basically correspond to $0 \leq \beta \leq +\infty, -\infty \leq \gamma \leq 0$, but $\beta + \gamma \leq 0$.

5. The subregion for $\beta \leq 0, \gamma \geq 0, \beta + \gamma \geq 0$:

This subregion $PR_2R_0C_1P$ is bounded by the line segments $\overline{PR_2}$ and $\overline{R_2R_0}$ and the curve $C_{\beta=0}$.

In this case, by the monotonic character of μ_2, namely (6.20) and (6.40),

- any point lying **above** $C_{\gamma=0}$ must correspond to $\gamma > 0$
- any point lying **above** $C_{\beta=0}$ must correspond to $\beta < 0$

This shows that the **interior** of this subregion consists of points corresponding to $\beta < 0, \gamma > 0$ and $\beta + \gamma > 0$.

Moreover, by the statement (6.39), the line segment $\overline{PR_2}$ of this subregion, **with the exception of the point** P, consists of points corresponding to $\beta = -\infty, \gamma = +\infty$, but $\beta + \gamma \geq 0$.

Conclusively, the points of this subregion, **with the exception of the point** P, basically correspond to $-\infty \leq \beta \leq 0, 0 \leq \gamma \leq +\infty$, but $\beta + \gamma \geq 0$.

6. The subregion for $\beta \geq 0, \gamma \leq 0, \beta + \gamma \geq 0$:

This subregion $PR_1R_0B_1P$ is bounded by the line segment $\overline{R_1R_0}$, the curve $C_{\gamma=0}$ and the arc PR_1.

In this case, by the monotonic character of μ_2, namely (6.20) and (6.40),

- any point lying **below** $C_{\gamma=0}$ must correspond to $\gamma < 0$
- any point lying **below** $C_{\beta=0}$ must correspond to $\beta > 0$

This shows that the **interior** of this subregion consists of points corresponding to $\beta > 0, \gamma < 0$ and $\beta + \gamma > 0$.

Moreover, by the statement (6.38), the curve PR_1 of this subregion, **with the exception of the point** P, consists of points corresponding to $\beta = +\infty, \gamma = -\infty$, but $\beta + \gamma \geq 0$.

Conclusively, the points of this subregion, **with the exception of the point** P, basically correspond to $0 \leq \beta \leq +\infty, -\infty \leq \gamma \leq 0$, but $\beta + \gamma \geq 0$.

7. The subregion for $\beta \leq 0, \gamma \geq 0, \beta + \gamma = 0$:

This subregion is simply the line segment R_2R_0. At the point R_0, we have $\beta = \gamma = 0$.

In this case, by the monotonic character of μ_2, namely (6.20) and (6.40),

- any point lying **above** $C_{\gamma=0}$ (i.e. above the point R_0) must correspond to $\gamma > 0$
- any point lying **above** $C_{\beta=0}$ (i.e. above the point R_0) must correspond to $\beta < 0$

This shows that the **interior** of this subregion consists of points corresponding to $\beta < 0, \gamma > 0$ and $\beta + \gamma = 0$.

Moreover, by the statement (6.39), the point R_2 of this subregion consists of points corresponding to $\beta = -\infty, \gamma = +\infty$, but $\beta + \gamma = 0$.

Conclusively, the points of this subregion basically correspond to $-\infty \leq \beta \leq 0, 0 \leq \gamma \leq +\infty$, but $\beta + \gamma = 0$.

8. The subregion for $\beta \geq 0, \gamma \leq 0, \beta + \gamma = 0$:

This subregion is simply the line segment R_1R_0. At the point R_0, we have $\beta = \gamma = 0$.

In this case, by the monotonic character of μ_2, namely (6.20) and (6.40),

- any point lying **below** $C_{\gamma=0}$ (i.e. below the point R_0) must correspond to $\gamma < 0$
- any point lying **below** $C_{\beta=0}$ (i.e. below the point R_0) must correspond to $\beta > 0$

This shows that the **interior** of this subregion consists of points corresponding to $\beta > 0, \gamma > 0$ and $\beta + \gamma = 0$.

Moreover, by the statement (6.38), the point R_1 of this subregion consists of points corresponding to $\beta = +\infty, \gamma = -\infty$, but $\beta + \gamma = 0$.

Conclusively, the points of this subregion basically correspond to $0 \leq \beta \leq +\infty, -\infty \leq \gamma \leq 0$, but $\beta + \gamma = 0$.

This **completes** the discussions subjecting to the **signs of** β, γ and $\beta+\gamma$ in the aforesaid **8 subregions** of $\overline{R}_{\mu_1,\mu_2}$ of this step 4.

Step 5:

In this step, we shall discuss about the points O and P of the region $\overline{R}_{\mu_1,\mu_2}$, the discussions of which we have not carried out so far.

Obviously, the points O and P picture the cases of $\mu_1 = \mu_2 = 0$ and $\mu_1 = \mu_2 = 1$ respectively.

Now, by considering the system of equations (6.34), we conclude that for $\mu_1 = \mu_2$ we must have either of the following situations:

- $e^{\beta x+\gamma x^2} = 0$ for x being within $[0,1]$ with the **probable exception** of $x = 0$.
- $e^{\beta x+\gamma x^2} = +\infty$ for x being within $[0,1]$ with the **probable exception** of $x = 0$.

Moreover, we have already seen in the previous steps that

- $(\beta = -\infty, \gamma = +\infty)$ is the only case, if the point (μ_1, μ_2) lies on the **line segment** $\overline{OR_2P}$ described by the equation $\mu_2 = \mu_1$ **with the exception of the points** O **and** P.
- $(\beta = +\infty, \gamma = -\infty)$ is the only case, if the point (μ_1, μ_2) lies on the **parabolic arc** OR_1P described by the equation $\mu_2 = \mu_1^2$ **with the exception of the points** O **and** P.

Therefore, both $(\beta = -\infty, \gamma = +\infty)$ and $(\beta = +\infty, \gamma = -\infty)$ are **ruled out possible solutions** of both $e^{\beta x+\gamma x^2} = 0$ and $e^{\beta x+\gamma x^2} = +\infty$. So, we are left with the only possibilities $(\beta = -\infty, \gamma = -\infty)$ and $(\beta = +\infty, \gamma = +\infty)$ to fulfill both $e^{\beta x+\gamma x^2} = 0$ and $e^{\beta x+\gamma x^2} = +\infty$.

Here, the relations (6.35), (6.36) and (6.37), namely $\frac{\partial \mu_1}{\partial \beta} > 0$, $\frac{\partial \mu_2}{\partial \gamma} > 0$ and $\frac{\partial \mu_1}{\partial \gamma} = \frac{\partial \mu_2}{\partial \beta} > 0$ principally say the following:

- if both μ_1 and μ_2 acquire their **minimum** values, i.e. $\mu_1 = \mu_2 = 0$, then both β and γ must also acquire their **minimum** values, i.e. $\beta = \gamma = -\infty$, so as to make $e^{\beta x+\gamma x^2} = 0$ for x being within $[0,1]$ with the **probable exception** of $x = 0$.

- if both μ_1 and μ_2 acquire their **maximum** values, i.e. $\mu_1 = \mu_2 = 1$, then both β and γ must also acquire their **maximum** values, i.e. $\beta = \gamma = +\infty$, so as to make $e^{\beta x + \gamma x^2} = +\infty$ for x being within $[0, 1]$ with the **probable exception** of $x = 0$.

Hence, as the **concluding statement** of this step 5, we arrive that

- the point O corresponds to $(\beta = -\infty, \gamma = -\infty)$.
- the point P corresponds to $(\beta = +\infty, \gamma = +\infty)$.

Step 6:

Thus, we have seen that all the possible images of $(\beta, \gamma) \in \mathbb{R}^2$ with regard to the described relationship (6.34) are contained in the region R_{μ_1,μ_2}.

Additionally, **all** the boundaries (i.e. **boundary points**) of R_{μ_1,μ_2} are also covered by $(\beta = \pm\infty, \gamma = \pm\infty)$.

Conclusively, by putting $\mu_1^{(\beta,\gamma)} = \frac{\int_0^1 x e^{\beta x + \gamma x^2} dx}{\int_0^1 e^{\beta x + \gamma x^2} dx}$ and $\mu_2^{(\beta,\gamma)} = \frac{\int_0^1 x^2 e^{\beta x + \gamma x^2} dx}{\int_0^1 e^{\beta x + \gamma x^2} dx}$, the **co-domain of the vector function** $\left(\mu_1^{(\beta,\gamma)}, \mu_2^{(\beta,\gamma)}\right)^T$ (of two variables, i.e of (β, γ)) **spans** the entire region R_{μ_1,μ_2} **for all** $(\beta, \gamma) \in \mathbb{R}^2$.

In other words (i.e. conversely), for every $(\mu_1, \mu_2) \in R_{\mu_1,\mu_2}$, there exists an unique pair $(\beta, \gamma) \in \mathbb{R}^2$. This proves the **existence** of the solution of (6.34) (or equivalently of (11.57)).

Step 7:

Now, let us draw our attention to the discrete case, i.e. let now us take X for a **discrete** random variable. In this case, the proof of the aforesaid existence can be given by considering certain **basic characteristic properties** of the probability mass function of X, viz. $f_{X|\{d\}}(x_j) = \frac{e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}}$, $j \in \{1, 2, \ldots, N\}$ for $0 = x_1 < x_2 < \ldots < x_N = 1$ in a simpler manner. For a discrete X, the discussions are somewhat **similar**, especially for a **large** value of N. These characteristic properties are described as

1. If μ_1 tends to 0 from right infinitesimally (i.e. $\mu_1 \to 0+$), then μ_2 is **compelled** to tend infinitesimally close to 0, simply because of $\mu_1^2 < \mu_2 < \mu_1$. In that case, the probability element $f_{X|\{d\}}(x_1)$ tends **infinitesimally** close to 1 and thereby all other $N-1$ probability elements $f_{X|\{d\}}(x_2), f_{X|\{d\}}(x_3), \ldots, f_{X|\{d\}}(x_N)$ individually tend **infinitesimally** close to 0.

2. If μ_1 tends to 1 from left infinitesimally (i.e. $\mu_1 \to 1-$), then μ_2 is **compelled** to tend infinitesimally close to 1, simply because of $\mu_1^2 < \mu_2 < \mu_1$. In that case, the probability element $f_{X|\{d\}}(x_N)$ tends **infinitesimally** close to 1 and thereby all other $N-1$ probability elements $f_{X|\{d\}}(x_1), f_{X|\{d\}}(x_2), \ldots, f_{X|\{d\}}(x_{N-1})$ individually tend **infinitesimally** close to 0.

3. For **any** fixedly chosen $0 < \mu_1 < 1$, if μ_2 made to tend infinitesimally close to its **least upper bound** μ_1 from **left** (i.e. $\mu_2 \to \mu_1-$), then the **sum of the two extreme probability elements** $f_{X|\{d\}}(x_1) + f_{X|\{d\}}(x_N)$ tends infinitesimally close to 1 and all other probability elements $f_{X|\{d\}}(x_2), f_{X|\{d\}}(x_3), \ldots, f_{X|\{d\}}(x_{N-1})$ individually tend **infinitesimally** close to 0. This has been fully elaborated in the subsection 5.7.1 (of the chapter 5).

4. For **any** fixedly chosen $0 < \mu_1 < 1$, μ_2 **cannot be made** to tend infinitesimally to its **lower bound** μ_1^2 from **right** (i.e. $\mu_2 \to \mu_1^2+$ is **not** possible) for every given $N \in \mathbb{N}$ or for every $\mathcal{X}_X$. In that case, μ_2 has to be meaningfully chosen, otherwise the probability distribution of X will **not exist**. This very problem has been handled in full details in the subsection 5.7.2 (of the chapter 5).

   However, μ_2 always has a **greatest lower bound** $glim(\mu_2, \mu_1, \mathcal{X}_X)$ **greater or equal to** μ_1^2, above which (or equal to which) the value of μ_2 ensures the existence of the probability distribution of X.

   This lower bound $glim(\mu_2, \mu_1, \mathcal{X}_X)$ of μ_2 depends on μ_1 as well as on the support $\mathcal{X}_X$. Precisely, $\mu_2 \geq glim(\mu_2, \mu_1, \mathcal{X}_X)$ **must necessarily hold** in this case, such that $glim(\mu_2, \mu_1, \mathcal{X}_X) \geq \mu_1^2$.

Conclusively, in the discrete case, it is evidently clear that, all the possible pairs $(\beta, \gamma) \in \mathbb{R}^2$, **with the exception of the cases** that correspond to

$\mu_2 < glim(\mu_2, \mu_1, \mathcal{X}_X))$, **span** the entire region R_{μ_1,μ_2}. This **exception** takes place in **discrete** cases of X, when $glim(\mu_2, \mu_1, \mathcal{X}_X) > \mu_1^2$ happens to hold.

In other words (i.e. conversely), for every $(\mu_1, \mu_2) \in R_{\mu_1,\mu_2}$, **with the (above stated) exception of the cases of** $\mu_2 < glim(\mu_2, \mu_1, \mathcal{X}_X))$ **in the discrete cases of** X, there exists an unique pair $(\beta, \gamma) \in \mathbb{R}^2$. This proves the **existence** of the solution of (11.37) (or **equivalently** of (11.36)).

Whence, by summarizing both the discrete and the continuous cases of X, we conclude that the solution of (6.8) **exists** for every $(\mu_1, \mu_2) \in R_{\mu_1,\mu_2}$, **with the (above stated) exception of the discrete case of** X.

This ultimately proves the **existence** of the solution of the system of equations (4.2) for $m = 2$. □

6.3.11 Classification of types of probability distributions

The character analysis of probability distributions of standard types, namely of **constant**, **monotone** and **uni-extremal** types, necessitates the **exact classification** of these types. Precisely, the question is, how these types are **exactly determined** with subject to the predeterminately given first **two** moments of the probability distribution.

For our discussions in this regard, the variance σ_Y^2 [2] of the random variable Y is allowed to vary within it's variability range $0 < \sigma_Y^2 < (\mu_Y^{(1)} - a)(b - \mu_Y^{(1)})$ with respect to the **fixedly chosen** first moment $\mu_Y^{(1)}$ of the random variable Y. In fact, we shall show that the variability of σ_Y^2 is pictured by the inequality $0 < \sigma_{Y,U}^2 \leq \sigma_{Y,L}^2 < (\mu_Y^{(1)} - a)(b - \mu_Y^{(1)})$, such that $\sigma_{Y,U}^2$ and $\sigma_{Y,L}^2$ are the **marginal variances** of Y for uni-modal and bathtub-shaped nature of the probability distribution of Y respectively. Depending on the values of $\sigma_{Y,U}^2$ and $\sigma_{Y,L}^2$, the probability distribution of Y is uni-modal, bathtub-shaped, strictly monotonic increasing, strictly monotonic decreasing or constant.

For the sake of simplicity, our discussions in this regard shall be confined to the **continuous** cases of the random variable of Y (or equivalently, of the **continuous** cases of the random variable X). Nextly, we proceed to find the expressions of $\sigma_{Y,U}^2$ and $\sigma_{Y,L}^2$.

[2] **equivalently**, the second moment $\mu_Y^{(2)}$ could be chosen in place of σ_Y^2

6.3.12 Marginal variances of probability distributions

It has to be noted that the determination of (local) **maximum** and **minimum** points of a graphically represented probability distribution by means of differential calculus is only possible in continuous cases. Of course, the same rule can be applied in discrete cases as well, **provided** the value of N is **large enough**.

With subject to fixedly given range of variability $\mathcal{X}_Y = [a, b]$ and the first moment $\mu_Y^{(1)}$ of the random variable Y, we shall **primarily** focus on the evaluations of the following, the descriptions of which are given elaborately in the **subsection 6.3.13**):

- The **limiting** (or **marginal**) variance $\sigma_{Y,U}^2$, such that the probability distribution of Y shall be **uni-modal**, if $\sigma_Y^2 \leq \sigma_{Y,U}^2$
- The **limiting** (or **marginal**) variance $\sigma_{Y,L}^2$, such that the probability distribution of Y shall be **bathtub-shaped**, if $\sigma_Y^2 \geq \sigma_{Y,L}^2$

where $\sigma_Y^2 = \mu_Y^{(2)} - \left(\mu_Y^{(1)}\right)^2$ is the user given variance of Y.

Corresponding to the usual linear transformation $X = \frac{Y-a}{b-a}$, we have

- $\sigma_{X,U}^2 = \frac{\sigma_{Y,U}^2}{(b-a)^2}$ is the **limiting variance** for the **uni-modality** of the probability distribution of X.

  For uni-modality, $\sigma_{X,U}^2 \leq \sigma^2$.

- $\sigma_{X,L}^2 = \frac{\sigma_{Y,L}^2}{(b-a)^2}$ is the **limiting variance** for the **bathtub-shapeliness** of the probability distribution of X.

  For bathtub-shapeliness, $\sigma_{X,L}^2 \geq \sigma^2$.

In this subsection, our objective shall be to **derive the expressions** of $\sigma_{X,U}^2$ and $\sigma_{X,L}^2$ and **illustrate** them **graphically**.

In course of our present discussions, for the sake of **simplicity**, we shall principally stick to the **continuous case** of X.

So, we shall proceed to **analyze** the **uni-extremal nature** (either **uni-modality nature** or the **bathtub-shapeliness nature**) of the probability density function of X by means of $\sigma_{X,U}^2$ and $\sigma_{X,L}^2$.

For this, the study of the **extreme point** of the probability density function of X is of **absolute necessity** in the form of the following proposition:

Proposition 6.3.8. *The probability density function of X being, as usual, given by*

$$f_{X|\{d\}}(x) = \frac{e^{\beta x+\gamma x^2}}{\int\limits_0^1 e^{\beta t+\gamma t^2}dt} \tag{6.46}$$

*such that $d = (\mu_1, \mu_2)$, $\mathcal{X}_X(\{d\}) = [0, 1]$ and $\sigma^2 = \mu_2 - \mu_1^2$, then the **extreme value** of the probability density $f_{X|\{d\}}(x)$ at the point $x = \frac{-\beta}{2\gamma}$ is **maximum** or **minimum** according as $\gamma \lesseqgtr 0$.*

*Proof of the **proposition 6.3.8**.* Here, by

$$\begin{aligned} f'_{X|\{d\}}(x) &= f_{X|\{d\}}(x)(\beta + 2\gamma x) \text{ and} \\ f''_{X|\{d\}}(x) &= f'_{X|\{d\}}(x)(\beta + 2\gamma x) + 2\gamma f_{X|\{d\}}(x) \end{aligned} \tag{6.47}$$

and by keeping $f_{X|\{d\}}(x) > 0$ for every $x \in [0, 1]$ in mind, we can easily see the following:

- only at $x = \frac{-\beta}{2\gamma}$ the derivative $f'_{X|\{d\}}(x)$ vanishes, i.e. $f'_{X|\{d\}}\left(\frac{-\beta}{2\gamma}\right) = 0$
- $f''_{X|\{d\}}\left(\frac{-\beta}{2\gamma}\right) = 2\gamma f_{X|\{d\}}\left(\frac{-\beta}{2\gamma}\right) \gtreqless 0$ according as $\gamma \gtreqless 0$

which clearly shows that the probability density function $f_{X|\{d\}}(x)$ can have at most one extremal value at $x = \frac{-\beta}{2\gamma}$ and hence the probability density curve is uni- extremal, provided

- $\gamma \neq 0$
- $0 \leq \frac{-\beta}{2\gamma} \leq 1$

Therefore, irrespective of whether the extremal point $\frac{-\beta}{2\gamma}$ of the probability density curve $f_{X|\{d\}}(x)$ lies within the interval $[0, 1]$ or not, it is conclusively clear that the extreme value $f_{X|\{d\}}\left(\frac{-\beta}{2\gamma}\right)$ of $f_{X|\{d\}}(x)$ is maximum or minimum according as

$$\gamma \lesseqgtr 0 \tag{6.48}$$

□

As the next step, we shall give **two** statements pertaining to the **marginality** of the **uni-extremal nature** of the probability density curve $f_{X|\{d\}}(x)$. This uni-extremal nature is **understandably** controlled by the **position of the extremal point** $x = \frac{-\beta}{2\gamma}$ of the probability density curve $f_{X|\{d\}}(x)$.

Statement 6.3.4 (Marginal and pure uni-extremal nature for $\mu_1 \neq \frac{1}{2}$). *If $0 < \frac{-\beta}{2\gamma} < 1$, then the probability density curve $f_{X|\{d\}}(x)$ is **purely** of uni-extremal nature.*

*If **either** $\frac{-\beta}{2\gamma} = 0$ **or** $\frac{-\beta}{2\gamma} = 1$, then the uni-extremal nature of the probability density curve $f_{X|\{d\}}(x)$ is **marginal**.*

*That is, **either** $\frac{-\beta}{2\gamma} < 0$ **or** $\frac{-\beta}{2\gamma} > 1$ would mean that the probability density curve $f_{X|\{d\}}(x)$ is **not** uni-extremal **anymore**.*

Statement 6.3.5 (Role of β and γ for the marginal uni-extremal nature). *The **marginality** (or the **limiting** case) of the uni-extremal nature of any of the probability density curves (i.e. curve of either **uni-modal shaped** or **bathtub-shaped**) necessitates either of the following three conditions:*

- $x = 0 = \frac{-\beta}{2\gamma} \iff \beta = 0$
- $x = 1 = \frac{-\beta}{2\gamma} \iff \beta = -2\gamma$
- $x = \frac{1}{2} = \frac{-\beta}{2\gamma} \iff \beta = -\gamma$

Exactly at this point, we are in a position to show that both these two above statements shall be utilized for determining the **marginal variances** for both **uni-modal** and **bathtub** shaped curves for $\mu_1 \lesseqgtr \frac{1}{2}$.

So, we shall proceed to analyze the cases for $\mu_1 \lesseqgtr \frac{1}{2}$ one by one. However, these cases can be **conveniently** categorized into the two following cases:

- $\mu_1 = \frac{1}{2}$. The marginality of the uni-extremal nature of the density function $f_{X|\{d\}}(x)$ corresponds to the point $x = \frac{1}{2}$.
- $\mu_1 \lessgtr \frac{1}{2}$. The marginality of the uni-extremal nature of the density function $f_{X|\{d\}}(x)$ corresponds to the points $x \in \{0, 1\}$.

Proposition 6.3.9 (Marginally uni-extremal nature at $x = \frac{1}{2}$, i.e. $\beta = -\gamma$). *For $\mu_1 = \frac{1}{2}$, the probability density function $f_{X|\{d\}}(x)$ is either* ***symmetric uni-extremal*** *or* ***uniform****. The marginal variances are given by $\sigma^2_{X,L} = \sigma^2_{X,U} = \frac{1}{12}$.*

Proof of the ***proposition 6.3.9****.* For $\mu_1 = \frac{1}{2}$, let us consider the following two points:

- by the proposition 6.3.6 (of **lemma 7**), $\mu_1 = \frac{1}{2} \Leftrightarrow \beta = -\gamma$.
- by the proposition 6.3.7 (of **lemma 8**), the **necessary** and **sufficient** condition for the probability density curve for X to be **symmetric** is $\mu_1 = \frac{1}{2}$ or equivalently $\beta = -\gamma$.

Here, it is intuitively clear that the marginal variances for both **uni-modal** and **bathtub** shaped curves merge to a common marginal variance. Here, the limiting (common) variance is nothing different from the variance of the uniform distribution, where $\mu_1 = \frac{1}{2}$, $\sigma^2_{X,L} = \sigma^2_{X,U} = \frac{1}{12}$ (or equivalently, when $\beta = \gamma = 0$). The proof of this very assertion is rather trivial, but still we must give a **complete analysis** of this picture of $\mu_1 = \frac{1}{2}$.

By the proposition 6.3.8 of rather by (6.48), the density curve for X remains

- **symmetric uni-modal**, so long $\gamma = -\beta < 0$ holds
- **symmetric bathtub- shaped**, so long $\gamma = -\beta > 0$ holds

Again, by the proposition 6.3.1 (of **lemma 2**), the **second moment** μ_2 (or equivalently the **variance** $\sigma^2 = \mu_2 - \frac{1}{4}$) **strictly** monotonically increases with γ (of course for the fixed $\mu_1 = \frac{1}{2}$). This evidently proves the following:

- at $\mu_2 = \frac{1}{3}$, $\gamma = 0$ and the density curve of X is **neither** uni-modal **nor** bathtub shaped
- when $\mu_2 < \frac{1}{3}$, $\gamma < 0$ and the density curve of X is **uni-modal**
- when $\mu_2 > \frac{1}{3}$, $\gamma > 0$ and the density curve of X is **bathtub-shaped**

and this brings us to conclude that $\sigma^2_{X,L} = \sigma^2_{X,U} = \frac{1}{3} - \frac{1}{2^2} = \frac{1}{12}$ is the marginal variance of both uni- modal and the bathtub- shaped density curves of X. This completes the proof of the **proposition 6.3.9** (referring to $\mu_1 = \frac{1}{2}$). □

Now, let us come to the cases for $\mu_1 \neq \frac{1}{2}$. In that case, by the proposition 6.3.6 (of **lemma 7**), we must have $\beta \neq -\gamma$.

Again, as we have discussed, since the marginality of the uni-modal nature or the bathtub-shaped nature of the density curve $f_{X|\{d\}}(x)$ correspond to either $\beta = 0$ or $\beta = -2\gamma$, in either of these two cases, only $\gamma \neq 0$ can therefore be of interest. Thus, by keeping this in mind, the marginality of the uni-extremal cases for $\mu_1 \neq \frac{1}{2}$ are described in form of propositions as follows:

Proposition 6.3.10 (Marginally uni-extremal nature at $x = 0$, i.e. when $\beta = 0$). *Here,*

- $\mu_1 < \frac{1}{2}$ *($\Leftrightarrow \gamma < 0$) signifies that the* ***uni-modal*** *density curve $f_{X|\{d\}}(x)$ has its* ***maximal*** *point at $x = 0$.*
- $\mu_1 > \frac{1}{2}$ *($\Leftrightarrow \gamma > 0$) signifies that the* ***bathtub*** *shaped density curve $f_{X|\{d\}}(x)$ has its* ***minimal*** *point at $x = 0$.*

Proof of the ***proposition 6.3.10****.* Here, $\mu_1 = \frac{\int_0^1 xe^{\gamma x^2}dx}{\int_0^1 e^{\gamma x^2}dx} = \frac{e^\gamma - 1}{2\gamma \int_0^1 e^{\gamma x^2}dx}$ and for a given $\mu_1 \neq \frac{1}{2}$, we must have $\mu_1 \gtrless \frac{1}{2} \Leftrightarrow \gamma \gtrless 0$, simply because

$$\begin{aligned} &\int_0^1 e^{\gamma x}dx \gtrless \int_0^1 e^{\gamma x^2}dx \text{ according as } \gamma \gtrless 0 \\ \Leftrightarrow &\frac{e^\gamma - 1}{\gamma} \gtrless \int_0^1 e^{\gamma x^2}dx \text{ according as } \gamma \gtrless 0 \\ \Leftrightarrow &\mu_1 = \frac{e^\gamma - 1}{2\gamma \int_0^1 e^{\gamma x^2}dx} \gtrless \frac{1}{2} \text{ according as } \gamma \gtrless 0 \end{aligned} \tag{6.49}$$

But, by the proposition 6.3.4 (of **lemma 5**) the equation $\mu_1 = \frac{\int_0^1 xe^{\gamma x^2}dx}{\int_0^1 e^{\gamma x^2}dx}$ is **uniquely solvable** for γ for the fixedly given $\beta = 0$.

Thus, by the proposition 6.3.8, the very fact $\gamma \gtrless 0 \Leftrightarrow f''_{X|\{d\}}\left(\frac{-\beta}{2\gamma}\right) \gtrless 0$ leads us to the following conclusion:

- $\mu_1 > \frac{1}{2} \Leftrightarrow \gamma > 0$ means that the point $x = 0$ is the **minimal** point of the **bathtub-shaped** density curve.

- $\mu_1 < \frac{1}{2} \Leftrightarrow \gamma < 0$ means that the point $x = 0$ is the **maximal** point of the **uni-modal** density curve.

This completes the proof of the **proposition 6.3.10**. □

Proposition 6.3.11 (Marginally uni-extremal nature at $x = 1$, i.e. when $\beta = -2\gamma$). *Here,*

- $\mu_1 > \frac{1}{2}$ *($\Leftrightarrow \gamma < 0$) signifies that the* ***uni-modal*** *density curve $f_{X|\{d\}}(x)$ has its* ***maximal*** *point at $x = 1$.*

- $\mu_1 < \frac{1}{2}$ *($\Leftrightarrow \gamma > 0$) signifies that the* ***bathtub*** *shaped density curve $f_{X|\{d\}}(x)$ has its* ***minimal*** *point at $x = 1$.*

Proof of the ***proposition 6.3.11****.* Here, $\mu_1 = \frac{\int_0^1 x e^{-2\gamma x + \gamma x^2} dx}{\int_0^1 e^{-2\gamma x + \gamma x^2} dx} = 1 - \frac{e^{\gamma} - 1}{2\gamma \int_0^1 e^{\gamma x^2} dx} \Leftrightarrow$ $1 - \mu_1 = \frac{e^{\gamma} - 1}{2\gamma \int_0^1 e^{\gamma x^2} dx}$.

Analogously (as in the case of the proposition 6.3.10), we can easily show in this very case that $\mu_1 \lesseqgtr \frac{1}{2} \Leftrightarrow \gamma \gtreqless 0$.

Exactly by the analogous arguments as in the case of the **proposition 6.3.10**, we arrive at the following conclusion:

- $1 - \mu_1 > \frac{1}{2} \Leftrightarrow \mu_1 < \frac{1}{2} \Leftrightarrow \gamma > 0$ means that the point $x = 1$ is the **minimal** point of the **bathtub-shaped** density curve.

- $1 - \mu_1 < \frac{1}{2} \Leftrightarrow \mu_1 > \frac{1}{2} \Leftrightarrow \gamma < 0$ means that the point $x = 1$ is the **maximal** point of the **uni-modal** density curve.

This completes the proof of the **proposition 6.3.11**. □

Hence, **with subject to the above propositions 6.3.10 and 6.3.11**, let us proceed to give the working rules for computations of the marginal variances of uni-modal and bathtub-shaped probability densities one by one as follows:

Proposition 6.3.12 (Working rule for the computation of $\sigma^2_{X,U}$ for a uni-modal density function in case of $\mu_1 < \frac{1}{2}$). *For the* ***marginal uni-modal*** *density function corresponding to $\mu_1 < \frac{1}{2}$, the equation*

$$\mu_1 = \frac{e^{\gamma} - 1}{2\gamma \int\limits_0^1 e^{\gamma x^2} dx} \tag{6.50}$$

is solved ***uniquely*** *for γ and subsequently the probability density function of X is given by*

$$f_{X|\{d\}}(x) = \frac{e^{\gamma x^2}}{\int\limits_0^1 e^{\gamma t^2} dt}, \text{ for } 0 \leq x \leq 1 \text{ and } d = (\mu_1, \sigma^2_{X,U} + \mu_1^2) \tag{6.51}$$

and the marginal variance $\sigma^2_{X,U}$ (i.e. the ***least upper bound*** *of the variance of X) is given by*

$$\sigma^2_{X,U} = \frac{\int\limits_0^1 x^2 e^{\gamma x^2} dx}{\int\limits_0^1 e^{\gamma x^2} dx} - \mu_1^2 \tag{6.52}$$

The proof of the **proposition 6.3.12**, namely the very fact that $\sigma^2_{X,U}$ is **actually** the marginal variance for the **uni-modality** for $\mu_1 < \frac{1}{2}$, is followed by the **proposition 6.3.18** (of the **subsection 6.3.13**).

Remark 6.3.4. *Conclusively, if $\sigma^2 \leq \sigma^2_{X,U}$, then the probability density $f_{X|\{d\}}(x)$ is* ***uni-modal*** *and if $\sigma^2 > \sigma^2_{X,U}$, then $f_{X|\{d\}}(x)$ is either* ***strictly monotonic decreasing*** *or* ***bathtub-shaped****.*

Proposition 6.3.13 (Working rule for the computation of $\sigma^2_{X,U}$ for a uni-modal density function in case of $\mu_1 > \frac{1}{2}$). *For the* ***marginal uni-modal*** *density function corresponding to $\mu_1 > \frac{1}{2}$, the equation*

$$1 - \mu_1 = \frac{e^{\gamma} - 1}{2\gamma \int\limits_0^1 e^{\gamma x^2} dx} \tag{6.53}$$

is solved ***uniquely*** *for γ and subsequently the probability density function of X is given by*

$$f_{X|\{d\}}(x) = \frac{e^{-2\gamma x + \gamma x^2}}{\int\limits_0^1 e^{-2\gamma t + \gamma t^2} dt}, \text{ for } 0 \leq x \leq 1 \text{ and } d = (\mu_1, \sigma^2_{X,U} + \mu_1^2) \tag{6.54}$$

and the marginal variance $\sigma^2_{X,U}$ (i.e. the ***least upper bound*** *of the variance of X) is given by*

$$\sigma^2_{X,U} = \frac{\int\limits_0^1 x^2 e^{-2\gamma x + \gamma x^2} dx}{\int\limits_0^1 e^{-2\gamma x + \gamma x^2} dx} - \mu_1^2 = \frac{\int\limits_0^1 x^2 e^{\gamma x^2} dx}{\int\limits_0^1 e^{\gamma x^2} dx} - (1 - \mu_1)^2 \tag{6.55}$$

The proof of the **proposition 6.3.13**, namely the very fact that $\sigma^2_{X,U}$ is **actually** the marginal variance for the **uni-modality** for $\mu_1 > \frac{1}{2}$, is followed by the **proposition 6.3.19** (of the **subsection 6.3.13**).

Remark 6.3.5. *Conclusively, if $\sigma^2 \leq \sigma^2_{X,U}$, then the probability density $f_{X|\{d\}}(x)$ is* ***uni-modal*** *and if $\sigma^2 > \sigma^2_{X,U}$, then $f_{X|\{d\}}(x)$ is either* ***strictly monotonic increasing*** *or* ***bathtub-shaped****.*

Proposition 6.3.14 (Working rule for the computation of $\sigma^2_{X,L}$ for a bathtub-shaped density function in case of $\mu_1 < \frac{1}{2}$). *For the* ***marginal bathtub-shaped*** *density function corresponding to $\mu_1 < \frac{1}{2}$, the equation*

$$1 - \mu_1 = \frac{e^{\gamma} - 1}{2\gamma \int\limits_0^1 e^{\gamma x^2} dx} \tag{6.56}$$

is solved ***uniquely*** *for γ and subsequently the probability density function of X is given by*

$$f_{X|\{d\}}(x) = \frac{e^{-2\gamma x + \gamma x^2}}{\int\limits_0^1 e^{-2\gamma t + \gamma t^2} dt}, \text{ for } 0 \leq x \leq 1 \text{ and } d = (\mu_1, \sigma^2_{X,L} + \mu_1^2) \tag{6.57}$$

and the marginal variance $\sigma^2_{X,L}$ (i.e. the ***greatest lower bound*** *of the variance of X) is given by*

$$\sigma^2_{X,L} = \frac{\int\limits_0^1 x^2 e^{-2\gamma x + \gamma x^2} dx}{\int\limits_0^1 e^{-2\gamma x + \gamma x^2} dx} - \mu_1^2 = \frac{\int\limits_0^1 x^2 e^{\gamma x^2} dx}{\int\limits_0^1 e^{\gamma x^2} dx} - (1 - \mu_1)^2 \tag{6.58}$$

The proof of the **proposition 6.3.14**, namely the very fact that $\sigma^2_{X,L}$ is **actually** the marginal variance for the **bathtub-shapeliness** for $\mu_1 < \frac{1}{2}$, is followed by the **proposition 6.3.18** (of the **subsection 6.3.13**).

Remark 6.3.6. *Conclusively, if $\sigma^2 \geq \sigma^2_{X,L}$, then the probability density $f_{X|\{d\}}(x)$ is* ***bathtub-shaped*** *and if $\sigma^2 < \sigma^2_{X,L}$, then $f_{X|\{d\}}(x)$ is either* ***strictly monotonic decreasing*** *or* ***uni-modal****.*

Proposition 6.3.15 (Working rule for the computation of $\sigma^2_{X,L}$ for a bathtub-shaped density function in case of $\mu_1 > \frac{1}{2}$). *For the* ***marginal bathtub-shaped*** *density function corresponding to $\mu_1 > \frac{1}{2}$, the equation*

$$\mu_1 = \frac{e^{\gamma} - 1}{2\gamma \int\limits_0^1 e^{\gamma x^2} dx} \tag{6.59}$$

is solved ***uniquely*** *for γ and subsequently the probability density function of X is given by*

$$f_{X|\{d\}}(x) = \frac{e^{\gamma x^2}}{\int\limits_0^1 e^{\gamma t^2} dt}, \text{ for } 0 \leq x \leq 1 \text{ and } d = (\mu_1, \sigma^2_{X,L} + \mu_1^2) \tag{6.60}$$

and the marginal variance $\sigma^2_{X,L}$ (i.e. the ***greatest lower bound*** *of the variance of X) is given by*

$$\sigma^2_{X,L} = \frac{\int\limits_0^1 x^2 e^{\gamma x^2} dx}{\int\limits_0^1 e^{\gamma x^2} dx} - \mu_1^2 \tag{6.61}$$

The proof of the **proposition 6.3.15**, namely the very fact that $\sigma^2_{X,L}$ is **actually** the marginal variance for the **bathtub-shapeliness** for $\mu_1 > \frac{1}{2}$, is followed by the **proposition 6.3.19** (of the **subsection 6.3.13**).

Remark 6.3.7. *Conclusively, if $\sigma^2 \geq \sigma^2_{X,L}$, then the probability density $f_{X|\{d\}}(x)$ is* ***bathtub-shaped*** *and if $\sigma^2 < \sigma^2_{X,L}$, then $f_{X|\{d\}}(x)$ is either* ***strictly monotonic increasing*** *or* ***uni-modal****.*

6.3.13 Characteristic behavior of the extremal point

The **movement** of the extremal point $-\frac{\beta}{2\gamma}$ of the probability density curve $f_{X|\{d\}}(x) = \frac{e^{\beta x+\gamma x^2}}{\int_0^1 e^{\beta t+\gamma t^2}dt}$, $0 \leq x \leq 1$, $d = (\mu_1, \mu_2)$ plays a **predominant role** in the classification of probability density types. Exactly this is what we are going to discuss in this subsection and for this discussion, we shall **include** the cases, when $\gamma = 0$.

Trivially, the probability density curve $f_{X|\{d\}}(x)$ represents a **constant** probability distribution, if $\beta = \gamma = 0$ and this **does not** need any elaboration. So, we proceed to discuss the **nontrivial** cases, i.e. when β and γ are **not both simultaneously zero**.

In the last subsection, we have been discussing about the **marginal variances** $\sigma^2_{X,U}$ and $\sigma^2_{X,L}$. We also know that this **marginality** is precisely the case, when the extremal point $-\frac{\beta}{2\gamma}$ is either **0** or **1**.

Now, the following question legitimately arises: How does the **change** in σ^2 has the **control** over the movement of the extremal point $-\frac{\beta}{2\gamma}$ for any **fixedly chosen** μ_1, so that we can control the **monotonicity** or **uni-extremity** of the probability density function $f_{X|\{d\}}(x)$? Exactly this is our main task.

So, if σ^2 ($= \mu_2 - \mu_1^2$) is made to move from 0 (it's greatest lower bound) to $\mu_1(1-\mu_1)$ (it's least upper bound), then the following statements are **restated** for the sake of our **convenience**:

1. By the proposition 6.3.1 (of **lemma 2**) and by $\sigma^2 = \mu_2 - \mu_1^2$,

$$\left.\frac{d\mu_2}{d\gamma}\right|_{\mu_1 \text{ is fixed}} = \left.\frac{d\sigma^2}{d\gamma}\right|_{\mu_1 \text{is fixed}} > 0 \tag{6.62}$$

   which evidently pictures that γ **increases** with the **increase** in σ^2.

2. By the proposition 6.3.3 (of **lemma 4**)

$$\left.\frac{d\beta}{d\gamma}\right|_{\mu_1 \text{ is fixed}} < 0 \tag{6.63}$$

   which evidently pictures that β **decreases** with the **increase** in γ and **vice versa**.

3. By (6.40) belonging to the **step 1** of the **theorem 6.3.1**

$$\left.\frac{d\mu_2}{d\beta}\right|_{\mu_1 \text{ is fixed}} = \left.\frac{d\sigma^2}{d\beta}\right|_{\mu_1 \text{ is fixed}} < 0 \tag{6.64}$$

which evidently pictures that β **decreases** with the **increase** in σ^2.

4. By (6.38) and (6.39) belonging to the **step 1** of the **theorem 6.3.1**,

$$\sigma^2 \to 0+ \Longleftrightarrow (\gamma \to -\infty \,\&\, \beta \to +\infty) \tag{6.65}$$

and

$$\sigma^2 \to \mu_1(1-\mu_1)- \Longleftrightarrow (\gamma \to +\infty \,\&\, \beta \to -\infty) \tag{6.66}$$

5. With reference to the **step 2** of the **theorem 6.3.1**,

$$\beta + \gamma \lessgtr 0 \text{ according as } \mu_1 \lessgtr \frac{1}{2} \tag{6.67}$$

6. If σ^2 is made to tend to 0 from **right** (i.e. if $\sigma^2 \to 0+$) for a **fixedly** chosen μ_1, both the probability densities $\frac{e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}{\int_0^1 e^{-\frac{(t-\mu_1)^2}{2\sigma^2}}dt}$, $0 \le x \le 1$ and $\frac{e^{\beta x+\gamma x^2}}{\int_0^1 e^{\beta t+\gamma t^2}dt}$, $0 \le x \le 1$ **converge** to a **limiting** probability distribution that is nothing different from the **discrete degenerated probability distribution** about the point μ_1.

   In this regard, if $\sigma^2 \to 0+$, then by (6.65), $\beta \to +\infty$ and $\gamma \to -\infty$, but we have

$$-\frac{\beta}{2\gamma} \to \mu_1 \tag{6.68}$$

   However, if $\sigma \to \mu_1(1-\mu_1)-$, then by (6.66), we have $\beta \to -\infty$ and $\gamma \to +\infty$, but **limiting value** of $-\frac{\beta}{2\gamma}$ shall **not** be investigated for this dissertation. For the sake of definiteness, let us assume $-\frac{\beta}{2\gamma} \to \mu_{1,1}$.

With this, let us arrive at two important propositions with regard to $\mu_1 \neq \frac{1}{2}$, the formal proofs of which are **rather difficult** at the moment. So, only the **proof-ideas** instead of formal rigorous proofs are given.

Proposition 6.3.16 (Behavior of the abscissa of the extremal point for $\mu_1 < \frac{1}{2}$). *If σ^2 is made to move from 0 to $\mu_1(1-\mu_1)$, then the abscissa $-\frac{\beta}{2\gamma}$ of the extremal point moves from **right** to **left**, provided **either** $\gamma < 0$ **or** $\gamma > 0$ is strictly maintained.*

*The **proof-idea** of the **proposition 6.3.16**.* Having $-\frac{\beta}{2\gamma} = -\frac{\beta+\gamma}{2\gamma} + \frac{1}{2}$, by (6.67), we have $-(\beta+\gamma) > 0$.

Now, by (6.63), β **increases** with the **decrease** in γ and vice versa and in fact, the magnitude of β **increases** with the **increase** in the magnitude of γ, if they (β and γ) are of **opposite signs**.

So, the magnitude of the **positive sum** $-(\beta+\gamma)$ **does not** change rapidly with the increase in σ^2.

But, since by (6.62), γ **increases** with the **increase** in σ^2 and since $\frac{1}{\gamma}$ is a decreasing function because of $\frac{d}{d\gamma}(\frac{1}{\gamma}) = -\frac{1}{\gamma^2} < 0$, it is well **intuitively assertible** that $\frac{-(\beta+\gamma)}{2\gamma}$ is a **decreasing function**.

Therefore, $-\frac{\beta}{2\gamma} = -\frac{\beta+\gamma}{2\gamma} + \frac{1}{2}$ is a **decreasing function** of σ^2 and hence $-\frac{\beta}{2\gamma}$ moves from **right** to **left** with the increase in σ^2.

This completes the proof-idea of the **proposition 6.3.16**. □

Proposition 6.3.17 (Behavior of the abscissa of the extremal point for $\mu_1 > \frac{1}{2}$). *If σ^2 is made to move from 0 to $\mu_1(1-\mu_1)$, then the abscissa $-\frac{\beta}{2\gamma}$ of the extremal point moves from **left** to **right**, provided **either** $\gamma < 0$ **or** $\gamma > 0$ is strictly maintained.*

*The **proof-idea** of the **proposition 6.3.17**.* In this case, by (6.67), we have $-(\beta+\gamma) < 0$.

Exactly in the **similar** manner as the proposition 6.3.16 has been sketched, $-\frac{\beta}{2\gamma} = -\frac{\beta+\gamma}{2\gamma} + \frac{1}{2}$ is a **increasing function** of σ^2 and hence $-\frac{\beta}{2\gamma}$ moves from **left** to **right** with the increase in σ^2

This completes the proof-idea of the **proposition 6.3.17**. □

Remark 6.3.8 (The case of $\gamma = 0$). *The reader of this dissertation may legitimately ask, why the case of $\gamma = 0$ is not handled in the above propositions 6.3.16 and 6.3.17. The question is precisely about the extremal point $-\frac{\beta}{2\gamma}$.*

*Before we go ahead, we would like to make another thing importantly clear: The case of $\beta = \gamma = 0$ is completely **ruled out**, otherwise $\mu_1 = \frac{1}{2}$ would be the case, which **does not** correspond to the propositions 6.3.16 and 6.3.17. Thus, $\beta \neq 0$ shall be the case here.*

So, the question arises, whether $-\frac{\beta}{2\gamma} = +\infty$ or $-\frac{\beta}{2\gamma} = -\infty$. Logically, this depends on the following:

- *The sign of β (since $\beta \neq 0$).*
- *Whether γ is made to approach 0 from **left** or from **right**.*

Thus, we arrive at the following statements:

- *If $\mu_1 < \frac{1}{2}$ then by (6.7), $-\beta > 0$.*

  Here, if $\gamma \to 0-$, then $-\frac{\beta}{2\gamma} = -\infty$ and if $\gamma \to 0+$, then $-\frac{\beta}{2\gamma} = +\infty$.

- *If $\mu_1 > \frac{1}{2}$ then by (6.7), $-\beta < 0$.*

  Here, if $\gamma \to 0-$, then $-\frac{\beta}{2\gamma} = +\infty$ and if $\gamma \to 0+$, then $-\frac{\beta}{2\gamma} = -\infty$.

*In general, we know very well that $-\frac{\beta}{2\gamma} = \pm\infty$ says that the probability distribution of X is a **minimum information** monotonic probability distribution.*

Remark 6.3.9 (The case of $\gamma = -\infty$). *In this case, the probability distribution of X is simply a **degenerated** probability distribution and this special case shall be discussed in this very chapter. This special case is ascribed to the attainment of σ^2 its **greatest lower bound**, namely $\sigma^2 = 0$.*

Remark 6.3.10 (The case of $\gamma = +\infty$). *In this case, the probability distribution of X is simply a **bernoulli** probability distribution and this special case shall be discussed in this very chapter. This special case is ascribed to the attainment of σ^2 its **least upper bound**, namely $\sigma^2 = \mu_1(1 - \mu_1)$.*

With this, after **excluding** the **trivial cases** of $\mu_1 \in \{0, 1\}$, we shall prove another set of propositions by keeping principally the following **restatements** in mind:

- With the **increase** in σ^2 from 0 to $\mu_1(1-\mu_1)$, γ **increases** (stated by (6.62)) and in fact, γ moves from $-\infty$ to $+\infty$ (referred to both (6.65) and (6.66)).
- With the **increase** in σ^2 from 0 to $\mu_1(1-\mu_1)$, β **decreases** (stated by (6.64)) and in fact, β moves from $+\infty$ to $-\infty$ (referred to both (6.65) and (6.66)).

Proposition 6.3.18 (For any fixed $\mu_1 < \frac{1}{2}$, the probability densities are classified into uni-extremal and monotonic decreasing types).

Proof of the ***proposition 6.3.18****.* In the course of change of σ^2 from 0 to $\mu_1(1-\mu_1)$, the **changing behaviors** of β, γ and $-\frac{\beta}{2\gamma}$ are described in **four disjoint** subintervals of the interval $(0, \mu_1(1-\mu_1))$ individually:

1. The subinterval given by $0 \leq \sigma^2 \leq \sigma^2_{X,U}$: Correspondingly, within this subinterval, we have the following:

- $+\infty \geq \beta \geq 0$, where $\beta = 0$ corresponds to the left end point $x = 0$ of the support $[0, 1]$.
- $-\infty \leq \gamma \leq \gamma_0$, where by the **proposition 6.3.4** (of **lemma 5**), γ_0 is uniquely determinable by solving $\mu_1 = \frac{\int_0^1 x e^{\gamma_0 x^2} dx}{\int_0^1 e^{\gamma_0 x^2} dx}$ for the given $\beta = 0$.
- $\mu_1 \geq -\frac{\beta}{2\gamma} \geq 0$

which says that, at $\sigma^2 = \sigma^2_{X,U}$, we have $\beta = 0$, $\gamma = \gamma_0$ and $-\frac{\beta}{2\gamma} = 0$.

Here, $0 \leq -\frac{\beta}{2\gamma} \leq \mu_1 < 1$ means that the probability density must necessarily be **uni-extremal**.

Moreover, by the **proposition 6.3.10** referring to the **left end point** of the support $[0, 1]$, namely $x = 0$, $\mu_1 < \frac{1}{2} \Leftrightarrow \gamma_0 < 0$ means that the extremal point $-\frac{\beta}{2\gamma_0} = 0$ is the **maximal point** of the **uni-modal** probability density.

Again, by the **proposition 6.3.16**, any slight **increase** of the value of σ^2 **higher** than $\sigma^2_{X,U}$ would mean $-\frac{\beta}{2\gamma} < 0$ implying that the probability density is **not** uni-extremal anymore. This means nothing, but the very fact that the **uni-modality** is **retained**, if $\sigma^2 \leq \sigma^2_{X,U}$ and the **uni-modality is marginal**, if $\sigma^2 = \sigma^2_{X,U}$.

Conclusively, the probability density is **uni-modal** within this subinterval.

2. The subinterval given by $\sigma^2_{X,U} < \sigma^2 \leq \sigma^2_{\beta_0}$: Correspondingly, within this subinterval, we have the following:

- $0 > \beta \geq \beta_0$, where by the **proposition 6.3.4** (of **lemma 5**), β_0 is uniquely determinable by solving $\mu_1 = \frac{\int_0^1 x e^{\beta_0 x} dx}{\int_0^1 e^{\beta_0 x} dx}$ for the given $\gamma = 0$, which is followed by $\sigma^2_{\beta_0} = \frac{\int_0^1 x^2 e^{\beta_0 x} dx}{\int_0^1 e^{\beta_0 x} dx} - \mu_1^2$.

  Obviously, $\beta_0 < 0$ and this gets reconfirmed by (6.7), namely

  $\mu_1 < \frac{1}{2} \Leftrightarrow \beta_0 < 0$.

- $\gamma_0 < \gamma \leq 0$. Obviously, $\gamma_0 < 0$ and γ may attain the value 0 **from left** only.

- $0 > -\frac{\beta}{2\gamma} \geq -\infty$

which says that, at $\sigma^2 = \sigma^2_{\beta_0}$, we have $\beta = \beta_0$, $\gamma = 0$ and $-\frac{\beta}{2\gamma} = -\infty$ (**Notably**, $-\frac{\beta}{2\gamma} = -\infty$ at $\sigma^2 = \sigma^2_{\beta_0}$ is well explained by the very fact that $\beta = \beta_0 < 0$ & by taking $\gamma \to 0-$).

Here, by taking the values of γ and β, such that $\gamma_0 < \gamma < 0$ and $0 > \beta > \beta_0$ and thereby $0 > -\frac{\beta}{2\gamma} > -\infty$, by using the **proposition 6.3.8** for $\gamma < 0$, it is clear that the extremal point $-\frac{\beta}{2\gamma}$ is the **maximal point** lying on the **left** of $x = 0$. Evidently, the probability density is **monotonic decreasing**.

Moreover, the case of $\gamma = 0$, $\beta = \beta_0 < 0$ represents a **minimum information monotonic decreasing** probability density.

Conclusively, the probability density is **monotonic decreasing** within this subinterval.

3. The subinterval given by $\sigma^2_{\beta_0} < \sigma^2 < \sigma^2_{X,L}$: Correspondingly, within this subinterval, we have the following:

- $\beta_0 > \beta > -2\gamma_1$, where by the **proposition 6.3.4** (of **lemma 5**), γ_1 is uniquely determinable by solving $\mu_1 = \frac{\int_0^1 xe^{-2\gamma_1 x+\gamma_1 x^2}dx}{\int_0^1 e^{-2\gamma_1 x+\gamma_1 x^2}dx} \Leftrightarrow 1-\mu_1 = \frac{\int_0^1 xe^{\gamma_1 x^2}dx}{\int_0^1 e^{\gamma_1 x^2}dx}$, i.e. γ_1 is uniquely determinable by solving $1-\mu_1 = \frac{\int_0^1 xe^{\gamma_1 x^2}dx}{\int_0^1 e^{\gamma_1 x^2}dx}$, which is followed by $\sigma^2_{X,L} = \frac{\int_0^1 x^2e^{-2\gamma_1 x+\gamma_1 x^2}dx}{\int_0^1 e^{-2\gamma_1 x+\gamma_1 x^2}dx} - \mu_1^2$.

  Obviously, $\gamma_1 > 0$ and this gets reconfirmed by (6.49), namely

  $1-\mu_1 \lessgtr \frac{1}{2} \Leftrightarrow \gamma_1 \lessgtr 0$ ($\Leftrightarrow \mu_1 \gtrless \frac{1}{2}$).

- $0 < \gamma < \gamma_1$. Obviously, $\gamma_1 > 0$ and γ may attain the value 0 **from right** only.

- $+\infty > -\frac{\beta}{2\gamma} > 1$

which says that, at $\sigma^2 = \sigma^2_{X,L}$, we have $\beta = -2\gamma_1$, $\gamma = \gamma_1$ and $-\frac{\beta}{2\gamma} = 1$ (**Notably**, $-\frac{\beta}{2\gamma} = +\infty$ at $\sigma^2 = \sigma^2_{\beta_0}$ is well explained by the very fact that $\beta = \beta_0 < 0$ & by taking $\gamma \to 0+$).

Here, by taking the values of γ and β, such that $0 < \gamma < \gamma_1$ and $\beta_0 > \beta > -2\gamma_1$ and thereby $+\infty > -\frac{\beta}{2\gamma} > +1$, by using the **proposition 6.3.8** for $\gamma > 0$, it is clear that the extremal point $-\frac{\beta}{2\gamma}$ is the **minimal point** lying on the **right** of $x = 1$. Evidently, the probability density is **monotonic decreasing**.

Moreover, the case of $\gamma = 0$, $\beta = \beta_0 < 0$ represents a **minimum information monotonic decreasing** probability density.

Conclusively, the probability density is **monotonic decreasing** within this subinterval.

4. The subinterval given by $\sigma^2_{X,L} \leq \sigma^2 \leq \mu_1(1-\mu_1)$: Correspondingly, within this subinterval, we have the following:

- $-2\gamma_1 \geq \beta \geq -\infty$, where $\beta = -2\gamma_1$ corresponds to the right end point $x = 1$ of the support $[0, 1]$.

- $\gamma_1 \leq \gamma \leq +\infty$, where by the **proposition 6.3.4** (of **lemma 5**), γ_1 is uniquely determinable by solving $\mu_1 = \frac{\int_0^1 x e^{-2\gamma_1 x + \gamma_1 x^2} dx}{\int_0^1 e^{-2\gamma_1 x + \gamma_1 x^2} dx} \Leftrightarrow 1 - \mu_1 = \frac{\int_0^1 x e^{\gamma_1 x^2} dx}{\int_0^1 e^{\gamma_1 x^2} dx}$, i.e. γ_1 is uniquely determinable by solving $1 - \mu_1 = \frac{\int_0^1 x e^{\gamma_1 x^2} dx}{\int_0^1 e^{\gamma_1 x^2} dx}$, which is followed by $\sigma^2_{X,L} = \frac{\int_0^1 x^2 e^{-2\gamma_1 x + \gamma_1 x^2} dx}{\int_0^1 e^{-2\gamma_1 x + \gamma_1 x^2} dx} - \mu_1^2$.

  Obviously, $\gamma_1 > 0$ and this gets reconfirmed by (6.49), namely $1 - \mu_1 \lesseqgtr \frac{1}{2} \Leftrightarrow \gamma_1 \lesseqgtr 0$ ($\Leftrightarrow \mu_1 \gtreqless \frac{1}{2}$).

- $1 \geq -\frac{\beta}{2\gamma} \geq \mu_{1,1}$

which says that, at $\sigma^2 = \mu_1(1-\mu_1)$, we have $\beta = -\infty$, $\gamma = +\infty$ and $-\frac{\beta}{2\gamma} = \mu_{1,1}$.

Here, $0 < \mu_{1,1} \leq -\frac{\beta}{2\gamma} \leq 1$ means that the probability density must necessarily be **uni-extremal**.

Moreover, by the **proposition 6.3.11** referring to the **right end point** of the support $[0, 1]$, namely $x = 1$, $\mu_1 < \frac{1}{2} \Leftrightarrow \gamma_1 > 0$ means that the extremal point $-\frac{\beta}{2\gamma_1} = -\frac{(-2\gamma_1)}{2\gamma_1} = 1$ is the **minimal point** of the **bathtub-shaped** probability density.

Again, by the **proposition 6.3.16**, any slight **decrease** of the value of σ^2 **lower** than $\sigma^2_{X,L}$ would mean $-\frac{\beta}{2\gamma} > 1$ implying that the probability density is **not** uni-extremal anymore. This means nothing, but the very fact that the **bathtub-shapeliness** is **retained**, if $\sigma^2 \geq \sigma^2_{X,L}$ and the **bathtub-shapeliness is marginal**, if $\sigma^2 = \sigma^2_{X,L}$.

Conclusively, the probability density is **bathtub-shaped** within this subinterval.

This completes the proof of the **proposition 6.3.18**. □

Remark 6.3.11. *For $\mu_1 < \frac{1}{2}$, in course of the movement of σ^2 from 0 to $\mu_1(1-\mu_1)$, then the following are therefore the cases:*

- *If $\gamma < 0$, then $-\frac{\beta}{2\gamma}$ is made to move from μ_1 to $-\infty$.*
- *If $\gamma > 0$, then $-\frac{\beta}{2\gamma}$ is made to move from $+\infty$ to $\mu_{1,1}$.*

*In each of the above cases, $-\frac{\beta}{2\gamma}$ is made to move from **right** to **left**, as stated by the **proposition 6.3.16**.*

Proposition 6.3.19 (For any fixed $\mu_1 > \frac{1}{2}$, the probability densities are classified into uni-extremal and monotonic increasing types).

Proof of the ***proposition 6.3.19****.* In the course of change of σ^2 from 0 to $\mu_1(1-\mu_1)$, the **changing behaviors** of β, γ and $-\frac{\beta}{2\gamma}$ are described in **four disjoint** subintervals of the interval $(0, \mu_1(1-\mu_1))$ individually:

1. The subinterval given by $0 \leq \sigma^2 \leq \sigma^2_{X,U}$: Correspondingly, within this subinterval, we have the following:

- $+\infty \geq \beta \geq -2\gamma_1$, where $\beta = -2\gamma_1$ corresponds to the right end point $x = 1$ of the support $[0, 1]$.
- $-\infty \leq \gamma \leq \gamma_1$, where by the **proposition 6.3.4** (of **lemma 5**), γ_1 is uniquely determinable by solving $\mu_1 = \frac{\int_0^1 x e^{-2\gamma_1 x + \gamma_1 x^2} dx}{\int_0^1 e^{-2\gamma_1 x + \gamma_1 x^2} dx} \Leftrightarrow 1 - \mu_1 = \frac{\int_0^1 x e^{\gamma_1 x^2} dx}{\int_0^1 e^{\gamma_1 x^2} dx}$, i.e. γ_1 is uniquely determinable by solving $1 - \mu_1 = \frac{\int_0^1 x e^{\gamma_1 x^2} dx}{\int_0^1 e^{\gamma_1 x^2} dx}$, which is followed by $\sigma^2_{X,U} = \frac{\int_0^1 x^2 e^{-2\gamma_1 x + \gamma_1 x^2} dx}{\int_0^1 e^{-2\gamma_1 x + \gamma_1 x^2} dx} - \mu_1^2$.

  Obviously, $\gamma_1 < 0$ and this gets reconfirmed by (6.49), namely $1 - \mu_1 \lesseqgtr \frac{1}{2} \Leftrightarrow \gamma_1 \lesseqgtr 0$ ($\Leftrightarrow \mu_1 \gtreqless \frac{1}{2}$).
- $\mu_1 \leq -\frac{\beta}{2\gamma} \leq 1$

which says that, at $\sigma^2 = \sigma^2_{X,U}$, we have $\beta = -2\gamma_1$, $\gamma = \gamma_1$ and $-\frac{\beta}{2\gamma} = 1$.

Here, $0 < \mu_1 \leq -\frac{\beta}{2\gamma} \leq 1$ means that the probability density must necessarily be **uni-extremal**.

Moreover, by the **proposition 6.3.11** referring to the **right end point** of the support $[0, 1]$, namely $x = 1$, $\mu_1 > \frac{1}{2} \Leftrightarrow \gamma_1 < 0$ means that the extremal point $-\frac{\beta}{2\gamma_0} = 0$ is the **maximal point** of the **uni-modal** probability density.

Again, by the **proposition 6.3.17**, any slight **increase** of the value of σ^2 **higher** than $\sigma^2_{X,U}$ would mean $-\frac{\beta}{2\gamma} > 1$ implying that the probability density is **not** uni-extremal anymore. This means nothing, but the very fact

that the **uni-modality** is **retained**, if $\sigma^2 \leq \sigma^2_{X,U}$ and the **uni-modality is marginal**, if $\sigma^2 = \sigma^2_{X,U}$.

Conclusively, the probability density is **uni-modal** within this subinterval.

2. The subinterval given by $\sigma^2_{X,U} < \sigma^2 \leq \sigma^2_{\beta_0}$: Correspondingly, within this subinterval, we have the following:

- $-2\gamma_1 > \beta \geq \beta_0$, where by the **proposition 6.3.4** (of **lemma 5**), β_0 is uniquely determinable by solving $\mu_1 = \frac{\int_0^1 x e^{\beta_0 x} dx}{\int_0^1 e^{\beta_0 x} dx}$ for the given $\gamma = 0$, which is followed by $\sigma^2_{\beta_0} = \frac{\int_0^1 x^2 e^{\beta_0 x} dx}{\int_0^1 e^{\beta_0 x} dx} - \mu_1^2$.

  Obviously, $\beta_0 > 0$ and this gets reconfirmed by (6.7), namely

  $\mu_1 > \frac{1}{2} \Leftrightarrow \beta_0 > 0$.

  Moreover, by the **proposition 6.3.4** (of **lemma 5**), γ_1 is uniquely determinable by solving $\mu_1 = \frac{\int_0^1 x e^{-2\gamma_1 x + \gamma_1 x^2} dx}{\int_0^1 e^{-2\gamma_1 x + \gamma_1 x^2} dx} \Leftrightarrow 1 - \mu_1 = \frac{\int_0^1 x e^{\gamma_1 x^2} dx}{\int_0^1 e^{\gamma_1 x^2} dx}$, i.e. γ_1 is uniquely determinable by solving $1 - \mu_1 = \frac{\int_0^1 x e^{\gamma_1 x^2} dx}{\int_0^1 e^{\gamma_1 x^2} dx}$, which is followed by $\sigma^2_{X,U} = \frac{\int_0^1 x^2 e^{-2\gamma_1 x + \gamma_1 x^2} dx}{\int_0^1 e^{-2\gamma_1 x + \gamma_1 x^2} dx} - \mu_1^2$.

  Obviously, $\gamma_1 < 0$ and this gets reconfirmed by (6.49), namely

  $1 - \mu_1 \lesseqgtr \frac{1}{2} \Leftrightarrow \gamma_1 \lesseqgtr 0$ ($\Leftrightarrow \mu_1 \gtreqless \frac{1}{2}$).

- $\gamma_1 < \gamma \leq 0$. Obviously, $\gamma_1 < 0$ and γ may attain the value 0 **from left** only.

- $1 < -\frac{\beta}{2\gamma} \leq +\infty$

which says that, at $\sigma^2 = \sigma^2_{\beta_0}$, we have $\beta = \beta_0$, $\gamma = 0$ and $-\frac{\beta}{2\gamma} = +\infty$ (**Notably**, $-\frac{\beta}{2\gamma} = +\infty$ at $\sigma^2 = \sigma^2_{\beta_0}$ is well explained by the very fact that $\beta = \beta_0 > 0$ & by taking $\gamma \to 0-$).

Here, by taking the values of γ and β, such that $\gamma_1 < \gamma < 0$ and $-2\gamma_1 > \beta > \beta_0$ and thereby $1 < -\frac{\beta}{2\gamma} < +\infty$, by using the **proposition 6.3.8** for $\gamma < 0$, it is clear that the extremal point $-\frac{\beta}{2\gamma}$ is the **maximal point** lying on the **right** of $x = 1$. Evidently, the probability density is **monotonic increasing**.

Moreover, the case of $\gamma = 0$, $\beta = \beta_0 > 0$ represents a **minimum information monotonic increasing** probability density.

Conclusively, the probability density is **monotonic increasing** within this subinterval.

3. The subinterval given by $\sigma^2_{\beta_0} < \sigma^2 < \sigma^2_{X,L}$: Correspondingly, within this subinterval, we have the following:

- $\beta_0 > \beta > 0$, where by the **proposition 6.3.4** (of **lemma 5**), β_0 is uniquely determinable by solving $\mu_1 = \frac{\int_0^1 x e^{\beta_0 x} dx}{\int_0^1 e^{\beta_0 x} dx}$ for the given $\gamma = 0$, which is followed by $\sigma^2_{\beta_0} = \frac{\int_0^1 x^2 e^{\beta_0 x} dx}{\int_0^1 e^{\beta_0 x} dx} - \mu_1^2$.

  Obviously, $\beta_0 > 0$ and this gets reconfirmed by (6.7), namely

  $\mu_1 > \frac{1}{2} \Leftrightarrow \beta_0 > 0$.

- $0 < \gamma < \gamma_0$, where by the **proposition 6.3.4** (of **lemma 5**), γ_0 is uniquely determinable by solving $\mu_1 = \frac{\int_0^1 x e^{\gamma_0 x^2} dx}{\int_0^1 e^{\gamma_0 x^2} dx}$ for the given $\beta = 0$.

  Obviously, $\gamma_0 > 0$ and γ may attain the value 0 **from right** only.

- $-\infty < -\frac{\beta}{2\gamma} < 0$

which says that, at $\sigma^2 = \sigma^2_{X,L}$, we have $\beta = 0$, $\gamma = \gamma_0$ and $-\frac{\beta}{2\gamma} = 0$
(**Notably**, $-\frac{\beta}{2\gamma} = -\infty$ at $\sigma^2 = \sigma^2_{\beta_0}$ is well explained by the very fact that $\beta = \beta_0 > 0$ & by taking $\gamma \to 0+$).

Here, by taking the values of γ and β, such that $0 < \gamma < \gamma_0$ and $\beta_0 > \beta > 0$ and thereby $-\infty < -\frac{\beta}{2\gamma} < 0$, by using the **proposition 6.3.8** for $\gamma > 0$, it is clear that the extremal point $-\frac{\beta}{2\gamma}$ is the **minimal point** lying on the **left** of $x = 0$. Evidently, the probability density is **monotonic increasing**.

Moreover, the case of $\gamma = 0$, $\beta = \beta_0 > 0$ represents a **minimum information monotonic increasing** probability density.

Conclusively, the probability density is **monotonic increasing** within this subinterval.

4. The subinterval given by $\sigma^2_{X,L} \leq \sigma^2 \leq \mu_1(1 - \mu_1)$: Correspondingly, within this subinterval, we have the following:

- $0 \geq \beta \geq -\infty$, where $\beta = 0$ corresponds to the left end point $x = 0$ of the support $[0, 1]$.
- $\gamma_0 \leq \gamma \leq +\infty$
- $0 \leq -\frac{\beta}{2\gamma} \leq \mu_{1,1}$

which says that, at $\sigma^2 = \mu_1(1 - \mu_1)$, we have $\beta = -\infty$, $\gamma = +\infty$ and $-\frac{\beta}{2\gamma} = \mu_{1,1}$.

Here, $0 \leq -\frac{\beta}{2\gamma} \leq \mu_{1,1} < 1$ means that the probability density must necessarily be **uni-extremal**.

Moreover, by the **proposition 6.3.10** referring to the **left end point** of the support $[0, 1]$, namely $x = 0$, $\mu_1 > \frac{1}{2} \Leftrightarrow \gamma_0 > 0$ means that the extremal point $-\frac{\beta}{2\gamma_0} = 0$ is the **minimal point** of the **bathtub-shaped** probability density.

Again, by the **proposition 6.3.17**, any slight **decrease** of the value of σ^2 **lower** than $\sigma^2_{X,L}$ would mean $-\frac{\beta}{2\gamma} < 0$ implying that the probability density is **not** uni-extremal anymore. This means nothing, but the very fact that the **bathtub-shapeliness** is **retained**, if $\sigma^2 \geq \sigma^2_{X,L}$ and the **bathtub-shapeliness is marginal**, if $\sigma^2 = \sigma^2_{X,L}$.

Conclusively, the probability density is **bathtub-shaped** within this subinterval.

This completes the proof of the **proposition 6.3.19**. □

Remark 6.3.12. *For $\mu_1 > \frac{1}{2}$, in course of the movement of σ^2 from 0 to $\mu_1(1-\mu_1)$, then the following are therefore the cases:*

- *If $\gamma < 0$, then $-\frac{\beta}{2\gamma}$ is made to move from μ_1 to $+\infty$.*
- *If $\gamma > 0$, then $-\frac{\beta}{2\gamma}$ is made to move from $-\infty$ to $\mu_{1,1}$.*

*In each of the above cases, $-\frac{\beta}{2\gamma}$ is made to move from **left** to **right**, as stated by the **proposition 6.3.17**.*

Proposition 6.3.20 (For $\mu_1 = \frac{1}{2}$, the probability densities are classified into symmetric uni-extremal and constant types).

Proof of the ***proposition 6.3.20****.* By the **proposition 6.3.6** (of **lemma 7**), $\mu_1 = \frac{1}{2} \Leftrightarrow \beta + \gamma = 0$, i.e. $\beta = -\gamma$ is the case here and therefore the extremal point in this case is **no** different from $-\frac{\beta}{2\gamma} = \frac{1}{2}$.

This means, $-\frac{\beta}{2\gamma}$ **does not** change with σ^2.

So, the case of $\beta = -\gamma = 0$ for a particular value of σ^2 is necessarily the case that cannot to be left out of consideration. **Obviously**, the case of $\beta = \gamma = 0$ for $\mu_1 = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$ is **nothing different from** the representation of the constant probability distribution.

By keeping this in mind, the course of change of σ^2 from 0 to $\mu_1(1 - \mu_1) = \frac{1}{4}$, the **changing behaviors** of β and γ $(= -\beta)$ are described in **three disjoint** subintervals (one of which is a **singleton** subset) of the interval $(0, \frac{1}{4})$ individually:

1. The subinterval given by $0 \leq \sigma^2 < \sigma^2_{X,U} = \frac{1}{12}$: Correspondingly, within this subinterval, we have the following:

- $+\infty \geq \beta > 0$.
- $-\infty \leq \gamma\,(= -\beta) < 0$.
- $-\frac{\beta}{2\gamma} = \frac{1}{2}$

which says that, at $\sigma^2 = \sigma^2_{X,U} = \frac{1}{12}$, we must necessarily have $\beta = 0$, $\gamma = 0$.

Here, $0 < \mu_1 = \frac{1}{2} = -\frac{\beta}{2\gamma} < 1$ means that the probability density must necessarily be **uni-extremal**.

Moreover, by the **propositions 6.3.8 and 6.3.9** with regard to $\gamma < 0$, the extremal point $-\frac{\beta}{2\gamma} = \frac{1}{2}$ is the **maximal point** of the **uni-modal** probability density.

Again, any slight **increase** of the value of σ^2 **higher** than $\sigma^2_{X,U} = \frac{1}{12}$ would mean that the probability density is **not** uni-modal anymore. This means nothing, but the very fact that the **uni-modality** is **retained**, if $\sigma^2 < \sigma^2_{X,U}$ and the **uni-modality is marginally violated**, if $\sigma^2 = \sigma^2_{X,U}$ or $\gamma = 0$.

Conclusively, the probability density is **uni-modal** within this subinterval.

2. The singleton subset given by $\sigma^2 = \sigma^2_{X,U} = \sigma^2_{X,L} = \frac{1}{12}$: Correspondingly, we have the following:

- $\beta = 0$.
- $\gamma\,(= -\beta) = 0$.

which says that the probability is **constant** at this singleton point subset described by $\sigma^2 = \frac{1}{12}$.

3. The subinterval given by $\frac{1}{12} = \sigma^2_{X,L} < \sigma^2 \leq \frac{1}{4}$: Correspondingly, within this subinterval, we have the following:

- $0 > \beta \geq -\infty$.
- $0 < \gamma\,(= -\beta) \leq +\infty$.
- $-\frac{\beta}{2\gamma} = \frac{1}{2}$

which says that, at $\sigma^2 = \frac{1}{4}$, we have $\beta = -\infty$, $\gamma = +\infty$ and $-\frac{\beta}{2\gamma} = \frac{1}{2}$.

Here, $0 < \mu_1 = \frac{1}{2} = -\frac{\beta}{2\gamma} < 1$ means that the probability density must necessarily be **uni-extremal**.

Moreover, by the **propositions 6.3.8 and 6.3.9** with regard to $\gamma > 0$, the extremal point $-\frac{\beta}{2\gamma} = \frac{1}{2}$ is the **minimal point** of the **bathtub-shaped** probability density.

Again, any slight **decrease** of the value of σ^2 **lower** than $\sigma^2_{X,L} = \frac{1}{12}$ would mean that the probability density is **not** bathtub-shaped anymore. This means nothing, but the very fact that the **bathtub-shapeliness** is **retained**, if $\sigma^2 > \sigma^2_{X,L}$ and the **bathtub-shapeliness is marginally violated**, if $\sigma^2 = \sigma^2_{X,L}$ or $\gamma = 0$.

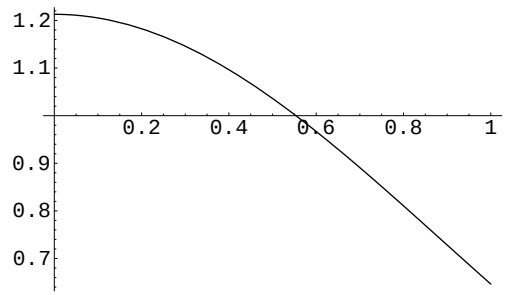
Conclusively, the probability density is **bathtub-shaped** within this subinterval.

This completes the proof of the **proposition 6.3.20**. □

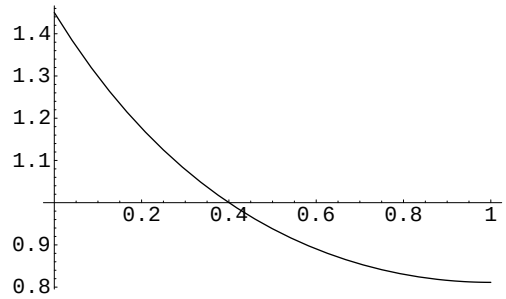
6.3.14 Graphical illustrations

Example 6.3.1 (An example for $\mu_1 = 0.45 < \frac{1}{2}$). *In that case, we must have*

- $\sigma^2_{X,U} = 0.0784466$ *and corresponding to* $d = (\mu_1, \sigma^2_{X,U} + \mu_1^2)$ *we have* $f_{X|\{d\}}(x) = 1.2129e^{-0.629682x^2}$
- $\sigma^2_{X,L} = 0.0850579$ *and corresponding to* $d = (\mu_1, \sigma^2_{X,L} + \mu_1^2)$ *we have* $f_{X|\{d\}}(x) = 1.4502e^{-1.16163x+0.580817x^2}$



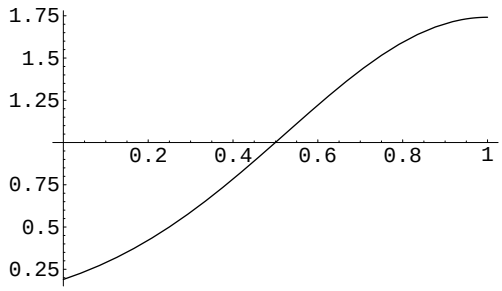
Uni-modal p.d.f. $1.2129e^{-0.629682x^2}$ for $\mu_1 = 0.45$



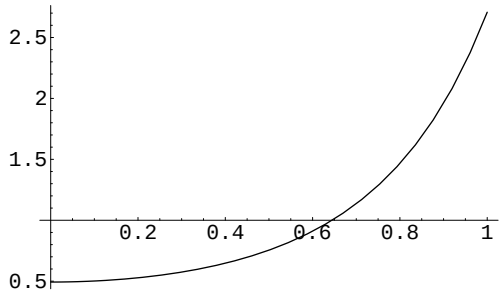
Bathtub-shaped p.d.f. $1.4502e^{-1.16163x+0.580817x^2}$ for $\mu_1 = 0.45$

Example 6.3.2 (An example for $\mu_1 = 0.65 > \frac{1}{2}$). *In that case, we must have*

- $\sigma^2_{X,U} = 0.0602757$ *and corresponding to* $d = (\mu_1, \sigma^2_{X,U} + \mu_1^2)$ *we have* $f_{X|\{d\}}(x) = 0.189806e^{4.43272x - 2.21636x^2}$
- $\sigma^2_{X,L} = 0.0786525$ *and corresponding to* $d = (\mu_1, \sigma^2_{X,L} + \mu_1^2)$ *we have* $f_{X|\{d\}}(x) = 0.492934e^{1.7033x^2}$



Uni-modal P.d.f. $0.189806e^{4.43272x - 2.21636x^2}$ for $\mu_1 = 0.65$



Bathtub-shaped P.d.f. $0.492934e^{1.7033x^2}$ for $\mu_1 = 0.65$

This completes our discussions about the determinations of the marginal variances of X. Just for the sake of complete clarity, let us come to the following summarized statement:

Statement 6.3.6 (The summary for the role of marginal variances). *Let us recapitulate the most important points that we have discussed in this very subsection*

- *For every $\mu_1 \in (0,1)$, we have $\mu_Y^{(1)} = a + (b-a)\mu_1 \in (a,b)$ and uniquely determined variances $\sigma^2_{X,U}$ and $\sigma^2_{X,L}$, such that $\sigma^2_{Y,U} = (b-a)^2\sigma^2_{X,U}$ and $\sigma^2_{Y,U} = (b-a)^2\sigma^2_{X,U}$*
- *For $\mu_1 \neq \frac{1}{2}$, $\sigma^2_{X,U} < \sigma^2_{X,L}$ and henceforth $\sigma^2_{Y,U} < \sigma^2_{Y,L}$, which corresponds to the case of uni- extremal distributions*
- *For $\mu_1 = \frac{1}{2}$, $\sigma^2_{X,U} = \sigma^2_{X,L} = \frac{1}{12}$ and henceforth $\sigma^2_{Y,U} = \sigma^2_{Y,L} = \frac{(b-a)^2}{12}$, which corresponds to the merged case of uniform distribution*
- *The uni-modal nature of the probability distribution* ***necessitates*** *$\gamma < 0$ and the bathtub-shapeliness nature of the probability distribution* ***necessitates*** *$\gamma > 0$.*

  However, ***none*** *of these conditions with regard to the sign of γ is* ***sufficient****.*
- *In case $\gamma \neq 0$ and at the same time, the probability distribution is not uni- extremal, then it is undoubtedly a monotonic probability distribution and in that case, $\frac{-\beta}{2\gamma} \notin (0,1)$.*

6.3.15 The monotonic character as special cases

For any given value of μ_1 lying strictly between 0 and 1, the probability density curve of X defined by of the probability density function denoted by $f_{X|\{d\}}(x) = \frac{e^{\beta x+\gamma x^2}}{\int_0^1 e^{\beta t+\gamma t^2}dt}$ for $d = (\mu_1, \mu_2)$ could be either **strictly monotonically increasing** or **strictly monotonically decreasing** under certain conditions.

This monotonicity takes place, when the variance $\sigma^2 = \mu_2 - \mu_1^2$ happens to lie strictly between $\sigma^2_{X,U}$ and $\sigma^2_{X,L}$, i.e. when

$$\sigma^2_{X,U} < \sigma^2 < \sigma^2_{X,L} \tag{6.69}$$

Our objective in this subsection shall be to examine, when the density curve is strictly monotonically increasing and when strictly monotonically decreasing.

For this, let state and prove the following proposition:

Proposition 6.3.21 (The monotonicity of the density function $f_{X|\{d\}}(x)$). *If the probability density function $f_{X|\{d\}}(x)$ happens to be monotonic, then the following hold:*

- $f_{X|\{d\}}(x)$ *is strictly monotonic* ***increasing****, if* $\mu_1 > \frac{1}{2}$
- $f_{X|\{d\}}(x)$ *is strictly monotonic* ***decreasing****, if* $\mu_1 < \frac{1}{2}$

Proof of the ***proposition 6.3.21****.* In order to prove the proposition, we shall discuss the four possible cases that arise.

Case 1: The density curve $f_{X|\{d\}}(x)$ is strictly monotonically **decreasing**, when

- the extremal point $\frac{-\beta}{2\gamma}$ happens to be the **left** of 0, i.e. $\frac{-\beta}{2\gamma} < 0$ and
- by (6.48), $\gamma < 0$ implies that the extremal point situated on the **left of 0** is **maximum**.

and therefore, by simultaneous consideration of $\frac{-\beta}{2\gamma} < 0$ and $\gamma < 0$, we get

$$\begin{aligned} -\beta > 0 &\Rightarrow \beta < 0 \\ &\Rightarrow \beta + \gamma < \gamma \\ &\Rightarrow \beta + \gamma < \gamma < 0 \ (\because \gamma < 0) \\ &\Rightarrow \beta + \gamma < 0 \\ &\Leftrightarrow \mu_1 < \frac{1}{2} \text{ (by the } \textbf{step 2} \text{ of the } \textbf{theorem 6.3.1} \text{)} \end{aligned} \tag{6.70}$$

Case 2: The density curve $f_{X|\{d\}}(x)$ is strictly monotonically **decreasing**, when

- the extremal point $\frac{-\beta}{2\gamma}$ happens to be the **right** of 1, i.e. $\frac{-\beta}{2\gamma} > 1$ and
- by (6.48), $\gamma > 0$ implies that the extremal point situated on the **right of 1** is **minimum**.

and therefore, by simultaneous consideration of $\frac{-\beta}{2\gamma} > 1$ and $\gamma > 0$, we get

$$\begin{aligned} -\beta > 2\gamma &\Rightarrow \beta < -2\gamma \\ &\Rightarrow \beta + \gamma < -\gamma \\ &\Rightarrow \beta + \gamma < -\gamma < 0 \ (\because \gamma > 0) \\ &\Rightarrow \beta + \gamma < 0 \\ &\Leftrightarrow \mu_1 < \frac{1}{2} \text{ (by the } \textbf{step 2} \text{ of the } \textbf{theorem 6.3.1} \text{)} \end{aligned} \tag{6.71}$$

Case 3: The density curve $f_{X|\{d\}}(x)$ is strictly monotonically **increasing**, when

- the extremal point $\frac{-\beta}{2\gamma}$ happens to be the **left** of 0, i.e. $\frac{-\beta}{2\gamma} < 0$ and
- by (6.48), $\gamma > 0$ implies that the extremal point situated on the **left of 0** is **minimum**.

and therefore, by simultaneous consideration of $\frac{-\beta}{2\gamma} < 0$ and $\gamma > 0$, we get

$$\begin{aligned} -\beta < 0 &\Rightarrow \beta > 0 \\ &\Rightarrow \beta + \gamma > \gamma \\ &\Rightarrow \beta + \gamma > \gamma > 0 \ (\because \gamma > 0) \\ &\Rightarrow \beta + \gamma > 0 \\ &\Leftrightarrow \mu_1 > \frac{1}{2} \text{ (by the } \textbf{step 2} \text{ of the } \textbf{theorem 6.3.1} \text{)} \end{aligned} \tag{6.72}$$

Case 4: The density curve $f_{X|\{d\}}(x)$ is strictly monotonically **increasing**, when

- the extremal point $\frac{-\beta}{2\gamma}$ happens to be the **right** of 1, i.e. $\frac{-\beta}{2\gamma} > 1$ and
- by (6.48), $\gamma < 0$ implies that the extremal point situated on the **right of 1** is **maximum**.

and therefore, by simultaneous consideration of $\frac{-\beta}{2\gamma} > 1$ and $\gamma < 0$, we get

$$\begin{aligned} -\beta < 2\gamma &\Rightarrow \beta > -2\gamma \\ &\Rightarrow \beta + \gamma > -\gamma \\ &\Rightarrow \beta + \gamma > -\gamma > 0 \ (\because \gamma < 0) \\ &\Rightarrow \beta + \gamma > 0 \\ &\Leftrightarrow \mu_1 > \frac{1}{2} \ (\text{ by the } \textbf{step 2} \text{ of the } \textbf{theorem 6.3.1} \ ) \end{aligned} \tag{6.73}$$

Conclusively, in cases, when (6.69) holds, i.e. if $\sigma^2_{X,U} < \sigma^2 < \sigma^2_{X,L}$, then the probability density curve of X is strictly monotonically **increasing** or **decreasing** according as

$$\mu_1 \gtrless \frac{1}{2} \tag{6.74}$$

and hence our **proposition 6.3.21** gets proved. □

6.3.16 Summary of probability density curves

The characteristics of the probability density curves of X can be briefly summarized with respect to the availability of moments one or two moments. In other words, assuming X to be continuous only, in monotone cases, the knowledge of μ_1 is necessary, whereas in uni- extremal cases, the knowledge of μ_1 and μ_2 is necessary.

It is a well known fact that, **in general**, any construction of a probability density function necessitates, at the very least, the following:

- the range of variability of the random variable
- a certain number $m \in \mathbb{N}_0$ of moments

As we have already mentioned, we shall however confine ourselves to the cases for $m \in \{0, 1, 2\}$, because we are principally interested in the cases of uniform, monotonic and uni- extremal probability distributions.

Without any loss of generality, we shall confine ourselves to the usage of the random variable X only (for the time being, we confine ourselves to the **continuous** X only), as **the random variable** Y **is basically nothing, but a linear transformation of** X. We trivially know that if no moments of probability distribution is available, i.e. $m = 0$, then the constructed probability distribution cannot be anything different from the well known uniform distribution.

In this subsection, we shall summarize the following cases that have been already discussed:

1. The range of variability of X being $[0, 1]$, how can the value of the first moment of X, namely μ_1, determine the type of the probability distribution ?
2. The range of variability of X being $[0, 1]$, how can the values of the first and the second moment of X, namely μ_1 and μ_2, determine the type of the probability distribution ?

Now, let us discuss the above two cases one by one:

Case 1:
Here, we must have $d = (\mu_1)$, thereby giving $f_{X|\{d\}}(x) = \frac{e^{\beta x}}{\int_0^1 e^{\beta t} dt}$ and thus

1. if $0 < \mu_1 < \frac{1}{2}$, then $\beta < 0$ and thus the probability distribution is **strictly monotonically decreasing**.
2. if $\mu_1 = \frac{1}{2}$, then $\beta = 0$ and thus the probability distribution is **uniform**.
3. if $\frac{1}{2} < \mu_1 < 1$, then $\beta > 0$ and thus the probability distribution is **strictly monotonically decreasing**.

Case 2:
Here, we must have $d = (\mu_1, \mu_2)$, thereby giving $f_{X|\{d\}}(x) = \frac{e^{\beta x + \gamma x^2}}{\int_0^1 e^{\beta t + \gamma t^2} dt}$ and thus

1. For $\mu_1 \neq \frac{1}{2}$,
   - if $\mu_1^2 < \mu_2 \leq \sigma_{X,U}^2 + \mu_1^2$, then the probability distribution is **uni-modal**.
   - if $\sigma_{X,U}^2 + \mu_1^2 < \mu_2 < \sigma_{X,L}^2 + \mu_1^2$, then the probability distribution is **monotonic** and in fact
     - if $\mu_1 < \frac{1}{2}$ then it is **strictly monotonically decreasing**
     - if $\mu_1 > \frac{1}{2}$ then it is **strictly monotonically increasing**
   - if $\sigma_{X,L}^2 + \mu_1^2 \leq \mu_2 < \mu_1$, then the probability distribution is **bathtub- shaped**.
2. For $\mu_1 = \frac{1}{2}$,
   - if $\frac{1}{4} < \mu_2 < \frac{1}{12} + \frac{1}{4} = \frac{1}{3}$, then the probability distribution is **symmetric uni- modal** and is symmetric about $x = \frac{1}{2}$.
   - if $\mu_2 = \frac{1}{3}$, then the probability distribution is **uniform**.
   - if $\frac{1}{12} + \frac{1}{4} = \frac{1}{3} < \mu_2 < \frac{1}{2}$, then the probability distribution is **symmetric bathtub- shaped** and is symmetric about $x = \frac{1}{2}$.

A Brief Note: Now, we shall come to the **discrete** case of X. If X happens to be a discrete random variable, then the same rules may be applied. These rules may be probably better applicable for large values of N. However, the rules cannot be derived exclusively with the help of derivatives, with the help of which the extrema values could be determined.

6.3.17 The characterizing expression

Till now, we have characterized the probability distributions of X with subject to the moments, mainly with regard to the probability density curves of a continuous X. **Trivially**, the characterization of the probability distributions of Y with subject to the moments means nothing more than a pure and simple **linear transformation** $X = \frac{Y-a}{b-a}$, such that $a = y_1$ and $b = y_N$ in case Y happens to be discrete for $\mathcal{X}_Y = \{y_1, y_2, \ldots, y_N\}$.

We know that

- the probability mass function or the probability density function of Y is denoted by $f_{Y|\{d_Y\}} = e^{\lambda_0+\lambda_1 y+\lambda_2 y^2}$, $y \in \mathcal{X}_Y$

- the probability mass function or the probability density function of $X = \frac{Y-a}{b-a}$ is denoted by $f_{X|\{d\}} = e^{\alpha+\beta x+\gamma x^2}$, $x \in \mathcal{X}_X$

Cases may arise, when the users, who use the software programs accepting the moments of the random variable $Y = a + (b-a)X$ as inputs and deliver the λ_0, λ_1 and λ_2 as outputs, wish to have the characterization of the probability distribution of Y with subject to the coefficients λ_0, λ_1 and λ_2.

In this subsection, we shall characterize the probability distributions of Y when its **first two moments are given as inputs**, with subject to the output values of λ_0, λ_1 and λ_2.

The function $f_{Y|\{d_Y\}}(y)$, $y \in \mathcal{X}_Y$, (with regard to $x = \frac{y-a}{b-a}$) is therefore rewritten in the following form, denoted as the **characterizing expression**, the **characterizing elements** being $\widetilde{\lambda}$ and $\widetilde{\mu}$ (here, notably, $K > 0$ and $\sigma_Y^2 = Var[Y] > 0$):

$$f_{Y|\{d_Y\}}(y) = Ae^{\frac{\beta}{b-a}(y-a)+\frac{\gamma}{(b-a)^2}(y-a)^2} = \frac{K}{\sigma_Y\sqrt{2\pi}}e^{\widetilde{\lambda}\left(\frac{y-\widetilde{\mu}}{\sigma_Y}\right)^2},\ y \in \mathcal{X}_Y \qquad (6.75)$$

such that

$$\frac{1}{A} = \begin{cases} \sum\limits_{j=1}^{N} e^{\beta x_j+\gamma x_j^2}, & : \ Y \text{ is discrete, such that } x_j = \frac{y_j-a}{b-a} \\ (b-a)\int\limits_0^1 e^{\beta x+\gamma x^2}dx & : \ Y \text{ is continuous, such that } x = \frac{y-a}{b-a} \end{cases} \qquad (6.76)$$

Notably,

- $f'_{Y|\{d_Y\}}(\widetilde{\mu}) = 0$
- $f''_{Y|\{d_Y\}}(\widetilde{\mu}) = \frac{2\widetilde{\lambda}K}{\sigma_Y^2\sqrt{2\pi}} \gtrless 0$ according as $\widetilde{\lambda} \gtrless 0$

Therefore, the utility of this very characterizing expression (6.75) is that it enables us to see the following immediately:

- if $\widetilde{\lambda} = 0$, then the probability distribution of Y is **uniform**.
- if $\widetilde{\lambda} < 0$, together with $\widetilde{\mu} < a$, then the probability distribution of Y is **strictly monotonically decreasing**.
- if $\widetilde{\lambda} > 0$, together with $\widetilde{\mu} > b$, then the probability distribution of Y is **strictly monotonically decreasing**.
- if $\widetilde{\lambda} < 0$, together with $\widetilde{\mu} > b$, then the probability distribution of Y is **strictly monotonically increasing**.
- if $\widetilde{\lambda} > 0$, together with $\widetilde{\mu} < a$, then the probability distribution of Y is **strictly monotonically increasing**.
- if $\widetilde{\lambda} < 0$, together with $a \leq \widetilde{\mu} \leq b$, then the probability distribution of Y is **uni- modal**.
- if $\widetilde{\lambda} > 0$, together with $a \leq \widetilde{\mu} \leq b$, then the probability distribution of Y is **bathtub- shaped**.

However, if the probability distribution is **multi-extremal**, then the above form has **absolutely no sense**.

Here, for to evaluate the values of $\widetilde{\lambda}$ and $\widetilde{\mu}$ in terms of λ_0, λ_1 and λ_2 so as to give the characterizing expression, we proceed as follows:

$$\begin{aligned}
& \gamma x^2 + \beta x = \frac{\gamma}{(b-a)^2}(y-a)^2 + \frac{\beta}{b-a}(y-a) \\
= \; & \frac{\gamma}{(b-a)^2}\left\{(y-a)^2 + \frac{\beta}{\gamma}(b-a)(y-a)\right\} \\
= \; & \frac{\gamma}{(b-a)^2}\left\{y^2 - 2ay + a^2 + \frac{\beta}{\gamma}(b-a)y - \frac{\beta}{\gamma}a(b-a)\right\} \\
= \; & \frac{\gamma}{(b-a)^2}\left\{y^2 - 2\left(a - \frac{\beta}{2\gamma}(b-a)\right)y + a^2 - \frac{\beta}{\gamma}a(b-a)\right\} \\
= \; & \frac{\gamma}{(b-a)^2}\left\{y - \left(a - \frac{\beta}{2\gamma}(b-a)\right)\right\}^2 - \frac{\beta^2}{4\gamma} \qquad (6.77)
\end{aligned}$$

Therefore, with the help of the following relation

$$f_{Y|\{d_Y\}}(y) = Ae^{\frac{\gamma}{(b-a)^2}\left[y-\left(a-\frac{\beta}{2\gamma}(b-a)\right)\right]^2 - \frac{\beta^2}{4\gamma}} = \frac{K}{\sigma_Y\sqrt{2\pi}}e^{\tilde{\lambda}\left(\frac{y-\tilde{\mu}}{\sigma}\right)^2} \qquad (6.78)$$

by comparing the coefficients of y in the expansion of the expressions of the power of e, we arrive at

- $\widetilde{\mu} = a - \frac{\beta}{2\gamma}(b-a) = \mu_Y^{(1)} + \delta_M$.

  If Y is continuous, then $\widetilde{\mu}$ is the extremum of the curve $f_{Y|\{d_Y\}}(y)$ and in that case, $\widetilde{\mu}$ is the mode of the probability distribution of Y, if $\widetilde{\lambda} < 0$. δ_M gives the difference between $\mu_Y^{(1)}$ and the extremum.

- $K = A\sigma_Y\sqrt{2\pi}e^{-\frac{\beta^2}{4\gamma}}$

- $\widetilde{\lambda} = \frac{\gamma\sigma_Y^2}{(b-a)^2}$

Finally, the values of λ_0, λ_1 and λ_2 are given as follows:

- $\lambda_0 = \log\left(\frac{K}{\sigma_Y\sqrt{2\pi}}\right) + \widetilde{\lambda}\frac{\widetilde{\mu}^2}{\sigma_Y^2}$
- $\lambda_1 = -2\widetilde{\lambda}\frac{\widetilde{\mu}}{\sigma_Y}$
- $\lambda_2 = \frac{\widetilde{\lambda}}{\sigma_Y^2}$

6.3.18 Probability distributions represented by boundary points defined by $(\mu_1, \mu_2) \in \partial \mathbf{D}^2$

Let us discuss the special cases referred to the uni-extremal probability distributions of X one by one

1. **Case for** $(\mu_2 = \mu_1^2, 0 < \mu_1 < 1) \Leftrightarrow (\beta = +\infty, \gamma = -\infty)$:

   $f_{X|\{(\mu_1,\mu_1^2)\}}(x)$ can be defined to represent a **discrete degenerated probability distribution** defined by $f_{X|\{(\mu_1,\mu_1^2)\}}(x) = 1$ for $x = \mu_1$

2. **Case for** $(\mu_2 = \mu_1, 0 < \mu_1 < 1) \Leftrightarrow (\beta = -\infty, \gamma = +\infty)$:

   $f_{X|\{(\mu_1,\mu_1)\}}(x)$ can be defined to represent a **bernoulli probability distribution** defined by

   $$f_{X|\{(\mu_1,\mu_1)\}}(x) = 1 - \mu_1 \text{ for } x = 0$$
   $$= \mu_1 \text{ for } x = 1$$

3. **Case for** $(\mu_2 = \mu_1 = 0) \Leftrightarrow (\beta = -\infty, \gamma = -\infty)$:

   $f_{X|\{(0,0)\}}(x)$ can be defined to represent a **discrete degenerated probability distribution** defined by $f_{X|\{(0,0)\}}(x) = 1$ for $x = 0$.

   This is the **special case of the bernoulli distribution** for $\mu_1 = 0$.

4. **Case for** $(\mu_2 = \mu_1 = 1) \Leftrightarrow (\beta = +\infty, \gamma = +\infty)$:

   $f_{X|\{(1,1)\}}(x)$ can be defined to represent a **discrete degenerated probability distribution** defined by $f_{X|\{(1,1)\}}(x) = 1$ for $x = 1$.

   This is the **special case of the bernoulli distribution** for $\mu_1 = 1$.

Notably, these special cases are basically referred to

- degenerated probability distributions (i.e. for $\mu_2 = \mu_1^2$), each of which have a **minimum variance**, i.e. $\sigma^2 = 0$
- bernoulli probability distributions (i.e. for $\mu_2 = \mu_1$), each of which have a **maximum variance**, i.e. $\sigma^2 = \mu_2 - \mu_1^2 = \mu_1(1 - \mu_1)$

This **fulfills the completeness** of our analysis of uni-extremal minimum information probability distributions.

6.3.19 Usages of uni-extremal probability distributions

The minimum information probability distributions did have two important usages in the field of the science of stochastics so far. In fact, as on date (March 2012), the existing limited company, named **Stochastikon GmbH situated in Würzburg**, GERMANY did make use of

- the software program developed by me, which computes λ_0, λ_1 and λ_2 values in **uni- extremal discrete cases** as well as
- the software program developed by me, which computes λ_0, λ_1 and λ_2 values in **uni- extremal continuous cases**

The development of both these above stated software programs, together with the development of the software programs for monotone cases is the **principle aim** of this dissertation. The two important usages of the uni-extremal case based software programs, as stated above, are hereby listed as follows:

1. the **continuous case** based software program for the computation of λ_0, λ_1 and λ_2 has been **intensively used** to develop the software package named **LEXPOL** and the development took place in the year 2006. The description of LEXPOL is referred to ([42]). This software package computes the **optimal prediction intervals** of maximum rotor load on the rotor blades of a wind turbine. This rotor load is due to the blast caused by the natural wind.

   This optimal prediction interval in each case is computable by the **repeated and multiple usages** of the continuous case based software program developed by me.

2. both the **discrete** and the **continuous case** based software programs (for the computation of λ_0, λ_1 and λ_2) have been used to develop further software programs in form of Java classes by me to compute **optimal prediction intervals** in cases when both the first and the second moment are estimated by interval estimation methods. That is, the interval estimations of both the first and the second moment are predetermined and given as inputs, after which optimal prediction intervals are given by my software programs as outputs.

Importantly, the utility of the software programs for uni- extremal cases in cases of both discrete and continuous lies principally on the very fact that

- the bell shape or the bathtub shape can be **shifted to the left or to the right** by adjusting the **first moment** carefully.

  The first moment, which is known to be the **position parameter**, is therefore can be used to **predetermine** the **position** of the bell shaped or of the bathtub shaped probability distribution curve.

- the **broadness** of the bell shapeliness or the bathtub shapeliness can be adjusted by adjusting the **variance** (or equivalently the **second moment**) carefully.

  The second moment, which is known to be the **shape parameter**, is therefore can be used to **predetermine** the **shape** (i.e the **thickness** or the **thinness**) of the bell shaped or of the bathtub shaped probability distribution curve.

Chapter 7

Standard parameters of standard m. i. probability distributions

As already mentioned, this dissertation aims primarily at developing a computer code for solving the problem of determination of an appropriate probability distribution $f_{Y|\{d_Y\}}(y)$, $y \in \mathcal{X}_Y(\{d_Y\})$ in cases, when the co-domain of the random structure function belongs to the constant, the minimum information monotone and the minimum information uni-extremal family respectively.

It may be however for important use to learn about the standard parameters of these probability distributions, notably the **quantile functions** that are rigorously used for **prediction procedures**.

A **feature size** of a probability distribution may be equivalently termed as a **probability function** or a **parameter**[1] of a probability distribution.

With regard to the uni-extremal cases, as discussed, I have developed the **software programs for the computations of λ_0, λ_1 and λ_2, both in discrete and continuous cases**. Further **software Java classes have been developed by me meant for computing the quantile functions**. The software Java classes for computing quantile functions together with the software Java classes meant for computing these aforesaid $\lambda_0, \lambda_1, \lambda_2$ val-

[1]The German word for **parameter** (of a **probability distribution**) could be given as **Kenngröße** (einer **Wahrscheinlichkeitsverteilung**)

ues are combined together to develop the **software Java classes to give prediction intervals as program outputs**.

In fact, the computed standard minimum information probability distributions do basically serve to determine the **situation based prediction intervals**.

However, the construction of constant probability distributions are rather trivial and the discussions regarding this are therefore avoided. Not only this, any programming work is too unnecessarily trivial. In fact, a constant probability distribution and a constant minimum information probability distribution have absolutely no difference.

In this chapter, in order to outline the relevant properties of the standard minimum information probability distributions, we shall briefly restate an important fact that the **boundedness** of the moments of Y is basically the result of the **compactness** of the range of variability of Y, namely $\mathscr{X}_Y(\{d_Y\})$.

Because of this compactness of the range of variability $\mathscr{X}_Y(\{d_Y\})$ of Y, special cases do arise, when the elements of $\mathscr{X}_Y(\{d_Y\})$ in case of a discrete Y are in an arithmetic progression with a common difference Δ. In that case, if $\mathscr{X}_Y(\{d_Y\}) = \{y_1, y_2, \ldots, y_N\}$, then

$$y_j = y_1 + (j-1)\Delta \text{ with } \Delta \in \mathbb{R} \setminus \{0\} \text{ and } j \in \{1, 2, \ldots, N\} \tag{7.1}$$

where Δ denotes the **increment** or the **decrement** according as Δ is positive or negative. With the help of this particular choice of $\mathscr{X}_Y(\{d_Y\})$ in a given situation, the essential feature sizes, namely the probability functions and the moments of special interests can be derived.

For the purpose of carrying out the derivations in cases when the discrete elements of $\mathcal{X}_Y(\{d_Y\})$ are in arithmetic progression as stated in (7.1), a linear transformation is introduced by an introduction of the random variable Z, such that $Y = y_1 + \Delta Z$. The following expressions are deduced with the help of solving certain difference equations:

$$E[Z] = \frac{1}{N}\sum_{k=1}^{N}(k-1) = \frac{1}{2}(N-1) \tag{7.2}$$

$$E[Z^2] = \frac{1}{N}\sum_{k=1}^{N}(k-1)^2 = \frac{1}{6}(N-1)(2N-1) \tag{7.3}$$

$$E[Z^3] = \frac{1}{N}\sum_{k=1}^{N}(k-1)^3 = \frac{1}{N}\left(\frac{N}{2}(N-1)\right)^2 \tag{7.4}$$

$$E[Z^4] = \frac{1}{N}\sum_{k=1}^{N}(k-1)^4 = \frac{1}{5}(N-1)\left(N^3 - \frac{3}{2}N^2 + \frac{1}{6}N + \frac{1}{6}\right) \tag{7.5}$$

In course of discussions in this chapter, there shall be certain **definitions** and **corollaries**, which are **too** trivial to be elaborated or be **formally proven**. In such cases, such elaborations or probably proofs are **unnecessary** and therefore **ignored**.

However, the important deductions of **sizable lengths** of this chapter are formally presented in form of **propositions**.

7.1 Monotone probability distribution

7.1.1 Discrete monotone probability distribution

As usual, we assume that the range of variability of $Y|\{d_Y\}$

$$\mathcal{X}_Y(\{d_Y\}) = \{y_1, y_2, \ldots, y_N\} \tag{7.6}$$

is known, the elements of which are arranged in the ascending order, i.e., $y_1 < y_2 < \ldots < y_N$. In this case, the minimum information probability distribution is, in general, an approximation. For obtaining it, besides the range of variability, the actual value $d_Y = (\mu_Y^{(1)})$ (i.e. the numerical value of the first moment) is additionally needed. The minimum information distribution is given by the following probability mass function of $Y|\{d_Y\}$:

$$f_{Y|\{d_Y\}}(y) = e^{\lambda_0 + \lambda_1 y} \text{ for } y \in \mathcal{X}_Y(\{d_Y\}) \tag{7.7}$$

with λ_0 and $\lambda_1 \neq 0$ uniquely determined by the solution of the following two equations:

$$\sum_{j=1}^{N} e^{\lambda_0 + \lambda_1 y_j} = 1 \tag{7.8}$$

$$\sum_{j=1}^{N} y_j e^{\lambda_0 + \lambda_1 y_j} = \mu_Y^{(1)} \tag{7.9}$$

If

- $\lambda_1 > 0$, then the probability distribution has a monotonic increasing probability mass function,
- $\lambda_1 < 0$, then the probability distribution has a monotonic decreasing probability mass function

The probability distribution (7.7) has the following characteristic properties:

Definition 7.1.1 (Probability mass function for $j = 1, 2, \ldots, N$).

$$f_{Y|\{d_Y\}}(y_j) = \frac{e^{\lambda_1 y_j}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}} \tag{7.10}$$

Definition 7.1.2 (Distribution and survival functions).

$$F_{Y|\{d_Y\}}(y_j) = \frac{\sum_{k=1}^{j} e^{\lambda_1 y_k}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}} \tag{7.11}$$

$$\overline{F}_{Y|\{d_Y\}}(y_j) = \frac{\sum_{k=j}^{N} e^{\lambda_1 y_k}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}} \tag{7.12}$$

Definition 7.1.3 (Quantile functions). *For $z \in (0, 1]$, the **upper** and the **lower** quantile functions are given as:*

$$Q^{(u)}_{Y|\{d_Y\}}(z) = y_{j_1} \tag{7.13}$$

such that j_1 is given by

- $\frac{\sum_{k=1}^{j_1-1} e^{\lambda_1 y_k}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}} < z \leq \frac{\sum_{k=1}^{j_1} e^{\lambda_1 y_k}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}}$

and

$$Q^{(\ell)}_{Y|\{d_Y\}}(z) = y_{j_2} \tag{7.14}$$

such that j_2 is given by

- $\frac{\sum_{k=j_2+1}^{N} e^{\lambda_1 y_k}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}} < z \leq \frac{\sum_{k=j_2}^{N} e^{\lambda_1 y_k}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}}$

Definition 7.1.4 (Failure intensity function).

$$a_{Y|\{d_Y\}}(y_j) = \frac{f_{Y|\{d_Y\}}(y_j)}{\overline{F}_{Y|\{d_Y\}}(y_j)} = \frac{e^{\lambda_1 y_j}}{\sum_{k=j}^{N} e^{\lambda_1 y_k}} \tag{7.15}$$

Definition 7.1.5 (Moments).

$$E[(Y|\{d_Y\})^i] = \frac{\sum_{k=1}^{N} y_k^i e^{\lambda_1 y_k}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}} \tag{7.16}$$

Definition 7.1.6 (Central moments).

$$\begin{aligned} E^c[(Y|\{d_Y\})^i] &= E\left[\Big(Y|\{d_Y\} - E[Y|\{d_Y\}]\Big)^i\right] \\ &= \frac{\sum\limits_{k=1}^{N}\Big(y_k - E[Y|\{d_Y\}]\Big)^i e^{\lambda_1 y_k}}{\sum\limits_{k=1}^{N} e^{\lambda_1 y_k}} \end{aligned} \tag{7.17}$$

Definition 7.1.7 (Standardized moments).

$$\begin{aligned} E^s[(Y|\{d_Y\})^i] &= E\left[\left(\frac{Y|\{d_Y\} - E[Y|\{d_Y\}]}{\sigma}\right)^i\right] \\ &= \frac{\sum\limits_{k=1}^{N}\left(\frac{y_k - E[Y|\{d_Y\}]}{\sigma}\right)^i e^{\lambda_1 y_k}}{\sum\limits_{k=1}^{N} e^{\lambda_1 y_k}} \end{aligned} \tag{7.18}$$

such that $\sigma = \sqrt{E^c[(Y|\{d_Y\})^2]}$

Definition 7.1.8 (Median).

$$Me[Y] = \frac{Q_Y^{(u)}(0.5) + Q_Y^{(\ell)}(0.5)}{2} \tag{7.19}$$

Definition 7.1.9 (Stochastic entropy).

$$H\left(P_{Y|\{d_Y\}}\right) = -\left(\lambda_0 + \lambda_1 \mu_Y^{(1)}\right) = -\left(\lambda_0 + \lambda_1 \frac{\sum\limits_{k=1}^{N} y_k e^{\lambda_1 y_k}}{\sum\limits_{k=1}^{N} e^{\lambda_1 y_k}}\right) \tag{7.20}$$

If the elements of the range of variability of $Y|\{d_Y\}$ follow an **arithmetic progression** described by (7.1), the probability functions are given as follows:

Statement 7.1.1 (Important summations).

$$\sum_{k=1}^{N} e^{\lambda_1 y_k} = \sum_{k=1}^{N} e^{\lambda_1 (y_1+(k-1)\Delta)} = e^{\lambda_1 y_1} \frac{1-e^{\lambda_1 \Delta N}}{1-e^{\lambda_1 \Delta}} \tag{7.21}$$

$$\sum_{k=1}^{j} e^{\lambda_1 y_k} = e^{\lambda_1 y_1} \frac{1-e^{\lambda_1 \Delta j}}{1-e^{\lambda_1 \Delta}} \tag{7.22}$$

$$\sum_{k=j+1}^{N} e^{\lambda_1 y_k} = e^{\lambda_1 y_1} \frac{1-e^{\lambda_1 \Delta N}-\left(1-e^{\lambda_1 \Delta j}\right)}{1-e^{\lambda_1 d}} = e^{\lambda_1 y_1} \frac{e^{\lambda_1 \Delta j}-e^{\lambda_1 \Delta N}}{1-e^{\lambda_1 \Delta}} \tag{7.23}$$

With this, we arrive at the following corollaries and propositions:

Corollary 7.1.1 (Probability mass function for $j = 1, 2, \ldots, N$).

$$f_{Y|\{d_Y\}}(y_j) = \frac{e^{\lambda_1 y_j}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}} = \frac{e^{\lambda_1 y_1} e^{\lambda_1 \Delta (j-1)}}{e^{\lambda_1 y_1} \frac{1-e^{\lambda_1 \Delta N}}{1-e^{\lambda_1 \Delta}}} = \frac{e^{\lambda_1 \Delta (j-1)} \left(1-e^{\lambda_1 \Delta}\right)}{1-e^{\lambda_1 \Delta N}} \tag{7.24}$$

Corollary 7.1.2 (Distribution and survival functions).

$$F_{Y|\{d_Y\}}(y_j) = \frac{\sum_{k=1}^{j} e^{\lambda_1 y_k}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}} = \frac{1-e^{\lambda_1 \Delta j}}{1-e^{\lambda_1 \Delta}} \frac{1-e^{\lambda_1 \Delta}}{1-e^{\lambda_1 \Delta N}} = \frac{1-e^{\lambda_1 \Delta j}}{1-e^{\lambda_1 \Delta N}} \tag{7.25}$$

$$\begin{aligned} \overline{F}_{Y|\{d_Y\}}(y_j) &= \frac{\sum_{k=j}^{N} e^{\lambda_1 y_k}}{\sum_{k=1}^{N} e^{\lambda_1 y_k}} = \frac{e^{\lambda_1 \Delta (j+1)} - e^{\lambda_1 \Delta N}}{1-e^{\lambda_1 \Delta}} \frac{1-e^{\lambda_1 \Delta}}{1-e^{\lambda_1 \Delta N}} \\ &= \frac{e^{\lambda_1 \Delta (j+1)} - e^{\lambda_1 \Delta N}}{1-e^{\lambda_1 \Delta N}} \end{aligned} \tag{7.26}$$

Corollary 7.1.3 (Quantile functions). *For every given $z \in (0,1]$, the **upper** and **lower** quantiles are given as*

$$Q^{(u)}_{Y|\{d_Y\}}(z) = y_{j_1} = y_1 + (j_1 - 1)\Delta \tag{7.27}$$

with subject to the inequality

$$\frac{1 - e^{\lambda_1 \Delta (j_1 - 1)}}{1 - e^{\lambda_1 \Delta N}} < z \leq \frac{1 - e^{\lambda_1 \Delta j_1}}{1 - e^{\lambda_1 \Delta N}} \tag{7.28}$$

and

$$Q^{(\ell)}_{Y|\{d_Y\}}(z) = y_{j_2} = y_1 + (j_2 - 1)\Delta \tag{7.29}$$

with subject to the inequality

$$\frac{e^{\lambda_1 \Delta j_2} - e^{\lambda_1 \Delta N}}{1 - e^{\lambda_1 \Delta N}} < z \leq \frac{e^{\lambda_1 \Delta (j_2 - 1)} - e^{\lambda_1 \Delta N}}{1 - e^{\lambda_1 \Delta N}} \tag{7.30}$$

As a matter of fact, the above two inequalities (7.30) and (7.28) follow directly by using (7.21), (7.22) and (7.23) on (7.14) and (7.13).

Corollary 7.1.4 (Failure intensity function).

$$\begin{aligned} a_{Y|\{d_Y\}}(y_j) &= \frac{e^{\lambda_1 y_j}}{\sum\limits_{k=j}^{N} e^{\lambda_1 y_k}} = e^{\lambda_1 \Delta (j-1)} \frac{1 - e^{\lambda_1 \Delta}}{e^{\lambda_1 \Delta (j+1)} - e^{\lambda_1 \Delta N}} \\ &= \frac{e^{\lambda_1 (j-1)\Delta} (1 - e^{\lambda_1 \Delta})}{e^{\lambda_1 \Delta (j+1)} - e^{\lambda_1 N}} \end{aligned} \tag{7.31}$$

Proposition 7.1.1 (Moments).

$$\begin{aligned} E[(Y|\{d_Y\})^i] &= \left(\frac{e^{\lambda_1 \Delta} - 1}{e^{\lambda_1 \Delta N} - 1} \right) \sum_{j=0}^{i} \binom{i}{j} y_1^{i-j} \Delta^j \left(\sum_{k=1}^{N} (k-1)^j e^{\lambda_1 \Delta (k-1)} \right) \\ &= \left(\frac{e^{\lambda_1 \Delta} - 1}{e^{\lambda_1 \Delta N} - 1} \right) \sum_{j=0}^{i} \binom{i}{j} y_1^{i-j} \Delta^j \, \frac{d^j}{d\xi^j} \left\{ \frac{e^{N\xi} - 1}{e^{\xi} - 1} \right\} \Big|_{\xi = \lambda_1 \Delta} \end{aligned} \tag{7.32}$$

Proof of the ***proposition 7.1.1****.*

$$
\begin{aligned}
E[(Y|\{d_Y\})^i] = \mu_Y^{(i)} &= \frac{\sum_{k=1}^{N} y_k^i e^{\lambda_1 \Delta (k-1)}}{\sum_{k=1}^{N} e^{\lambda_1 \Delta (k-1)}} \\
&= \frac{\sum_{k=1}^{N} \left(y_1 + (k-1)\Delta\right)^i e^{\lambda_1 \Delta (k-1)}}{\sum_{k=1}^{N} e^{\lambda_1 \Delta (k-1)}} \\
&= \frac{\sum_{k=1}^{N} \sum_{j=0}^{i} \binom{i}{j} y_1^{i-j} \left\{(k-1)\Delta\right\}^j e^{\lambda_1 \Delta (k-1)}}{\sum_{k=1}^{N} e^{\lambda_1 \Delta (k-1)}} \\
&= \left(\frac{1-e^{\lambda_1 \Delta}}{1-e^{\lambda_1 \Delta N}}\right) \sum_{j=0}^{i} \binom{i}{j} y_1^{i-j} \left(\sum_{k=1}^{N} \left\{(k-1)\Delta\right\}^j e^{\lambda_1 \Delta (k-1)}\right) \\
&= \left(\frac{e^{\lambda_1 \Delta}-1}{e^{\lambda_1 \Delta N}-1}\right) \sum_{j=0}^{i} \binom{i}{j} y_1^{i-j} \Delta^j \left(\sum_{k=1}^{N} (k-1)^j e^{\lambda_1 \Delta (k-1)}\right) \qquad (7.33) \\
&= \left(\frac{e^{\lambda_1 \Delta}-1}{e^{\lambda_1 \Delta N}-1}\right) \sum_{j=0}^{i} \binom{i}{j} y_1^{i-j} \Delta^j \, \frac{d}{d\xi} \left\{\sum_{k=1}^{N} (k-1)^{j-1} e^{(k-1)\xi}\right\}\Bigg|_{\xi=\lambda_1 \Delta} \\
&\quad \vdots \\
&= \left(\frac{e^{\lambda_1 \Delta}-1}{e^{\lambda_1 \Delta N}-1}\right) \sum_{j=0}^{i} \binom{i}{j} y_1^{i-j} \Delta^j \, \frac{d^j}{d\xi^j} \left\{\sum_{k=1}^{N} e^{(k-1)\xi}\right\}\Bigg|_{\xi=\lambda_1 \Delta} \\
&= \left(\frac{e^{\lambda_1 \Delta}-1}{e^{\lambda_1 \Delta N}-1}\right) \sum_{j=0}^{i} \binom{i}{j} y_1^{i-j} \Delta^j \, \frac{d^j}{d\xi^j} \left\{\frac{e^{N\xi}-1}{e^{\xi}-1}\right\}\Bigg|_{\xi=\lambda_1 \Delta} \qquad (7.34)
\end{aligned}
$$

and this completes the deduction proposed by the **proposition 7.1.1**. □

As a special case for $i = 1$,

Proposition 7.1.2 (The first moment).

$$
E[Y|\{d_Y\}] = \mu_Y^{(1)} = y_1 + \Delta \left(N - 1 + \frac{N}{e^{\lambda_1 \Delta N}-1} - \frac{1}{e^{\lambda_1 \Delta}-1}\right) \qquad (7.35)
$$

Proof of the ***proposition 7.1.2****.*

$$\begin{aligned}
E[Y|\{d_Y\}] &= \mu_Y^{(1)} = \tfrac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1}\left[y_1\tfrac{e^{\lambda_1\Delta N}-1}{e^{\lambda_1\Delta}-1}+\Delta\,\tfrac{d}{d\xi}\left(\tfrac{e^{\xi N}-1}{e^{\xi}-1}\right)\Big|_{\xi=\lambda_1\Delta}\right] \\
&= y_1+\Delta\,\tfrac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1}\left(\tfrac{\left(e^{\xi}-1\right)\left(Ne^{N\xi}\right)-\left(e^{N\xi}-1\right)e^{\xi}}{\left(e^{\xi}-1\right)^2}\right)\Big|_{\xi=\lambda_1\Delta} \\
&= y_1+\Delta\,\tfrac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1}\left(\tfrac{Ne^{N\xi}}{e^{\xi}-1}-\tfrac{\left(e^{N\xi}-1\right)e^{\xi}}{\left(e^{\xi}-1\right)^2}\right)\Big|_{\xi=\lambda_1\Delta} \qquad (7.36) \\
&= y_1+\Delta\,\frac{\left(e^{\lambda_1\Delta}-1\right)Ne^{\lambda_1\Delta N}-e^{\lambda_1\Delta}\left(e^{\lambda_1\Delta N}-1\right)}{\left(e^{\lambda_1\Delta}-1\right)\left(e^{\lambda_1\Delta N}-1\right)} \\
&= y_1+\Delta\,\left(\frac{Ne^{\lambda_1\Delta N}}{e^{\lambda_1\Delta N}-1}-\frac{e^{\lambda_1\Delta}}{e^{\lambda_1\Delta}-1}\right) \\
&= y_1+\Delta\,\left(\frac{N\left(e^{\lambda_1\Delta N}-1+1\right)}{e^{\lambda_1\Delta N}-1}-\frac{e^{\lambda_1\Delta}-1+1}{e^{\lambda_1\Delta}-1}\right) \\
&= y_1+\Delta\,\left(N-1+\frac{N}{e^{\lambda_1\Delta N}-1}-\frac{1}{e^{\lambda_1\Delta}-1}\right) \qquad (7.37)
\end{aligned}$$

and this completes the deduction proposed by the **proposition 7.1.2**. □

Corollary 7.1.5 (Central moments). *By setting* $Y|\{d_Y\} = y_1 + \Delta Z|\{d_Y\}$*, we write*

$$\begin{aligned}
E^c[(Y|\{d_Y\})^i] &= \mu_i^{(c)} = E\left[\left(Y|\{d_Y\}-\mu_Y^{(1)}\right)^i\right] \\
&= E\left[\left(y_1+\Delta Z|\{d_Y\}-y_1-\Delta E[Z|\{d_Y\}]\right)^i\right] \\
&= \Delta^i E^c[(Z|\{d_Y\})^i] \qquad (7.38)
\end{aligned}$$

where

$$E[(Z|\{d_Y\})^p] = \left(\frac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1}\right)\frac{d^p}{d\xi^p}\left\{\frac{e^{N\xi}-1}{e^{\xi}-1}\right\}\Big|_{\xi=\lambda_1\Delta} \qquad (7.39)$$

As a special case for $i = 2$, we arrive at the following proposition.

Proposition 7.1.3 (Variance (the second central moment)).

$$\begin{aligned}
& E^c\left[(Y|\{d_Y\})^2\right] = \sigma_Y^2 \\
= \;& \tfrac{\Delta^2}{\left(e^{\lambda_1\Delta}-1\right)^2\left(e^{\lambda_1\Delta N}-1\right)^2}\left\{e^{\lambda_1\Delta}\left(e^{\lambda_1\Delta N}-1\right)^2-N^2e^{\lambda_1\Delta N}\left(e^{\lambda_1\Delta}-1\right)^2\right\}
\end{aligned}$$

Proof of the ***proposition 7.1.3****.*

$$
\begin{aligned}
E^c[(Y|\{d_Y\})^2] = \mu_2^{(c)} = \sigma^2 &= \Delta^2 \left\{ E\left[Z|\{d_Y\}^2 \right] - (E[Z|\{d_Y\}])^2 \right\} \\
&= \Delta^2 \left\{ \frac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1} \frac{d^2}{d\xi^2} \left(\frac{e^{N\xi}-1}{e^{\xi}-1} \right) \Big|_{\xi=\lambda_1\Delta} - (E[Z|\{d_Y\}])^2 \right\} \\
&= \Delta^2 \left\{ \frac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1} \frac{d}{d\xi} \left(\frac{Ne^{N\xi}}{e^{\xi}-1} - \frac{e^{\xi}\left(e^{N\xi}-1\right)}{\left(e^{\xi}-1\right)^2} \right) \Big|_{\xi=\lambda_1\Delta} - (E[Z|\{d_Y\}])^2 \right\} \\
&= \Delta^2 \left\{ \frac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1} \frac{d}{d\xi} \left(\frac{(N-1)e^{N\xi}+1}{e^{\xi}-1} - \frac{e^{N\xi}-1}{\left(e^{\xi}-1\right)^2} \right) \Big|_{\xi=\lambda_1\Delta} - (E[Z|\{d_Y\}])^2 \right\} \\
&= \Delta^2 \left\{ \frac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1} \left(\frac{\left(e^{\xi}-1\right)\left(N(N-1)e^{N\xi}\right) - \left((N-1)e^{N\xi}+1\right)e^{\xi}}{\left(e^{\xi}-1\right)^2} \right. \right. \\
&\qquad \left. \left. - \frac{\left(e^{\xi}-1\right)^2\left(Ne^{N\xi}\right) - 2\left(e^{N\xi}-1\right)\left(e^{\xi}-1\right)e^{\xi}}{\left(e^{\xi}-1\right)^4} \right) \Big|_{\xi=\lambda_1\Delta} - (E[Z|\{d_Y\}])^2 \right\} \\
&= \Delta^2 \left\{ \frac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1} \left(\frac{N(N-1)e^{N\xi}}{e^{\xi}-1} - \frac{e^{\xi}\left((N-1)e^{N\xi}+1\right)}{\left(e^{\xi}-1\right)^2} - \frac{Ne^{N\xi}}{\left(e^{\xi}-1\right)^2} \right. \right. \\
&\qquad \left. \left. + \frac{2e^{\xi}\left(e^{N\xi}-1\right)}{\left(e^{\xi}-1\right)^3} \right) \Big|_{\xi=\lambda_1\Delta} - (E[Z|\{d_Y\}])^2 \right\} \\
&= \Delta^2 \left\{ \frac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1} \left(\frac{N(N-1)e^{N\xi}}{e^{\xi}-1} - \frac{\left(e^{\xi}-1+1\right)\left((N-1)e^{N\xi}+1\right)}{\left(e^{\xi}-1\right)^2} \right. \right. \\
&\qquad \left. \left. - \frac{Ne^{N\xi}}{\left(e^{\xi}-1\right)^2} + \frac{2\left(e^{\xi}-1+1\right)\left(e^{N\xi}-1\right)}{\left(e^{\xi}-1\right)^3} \right) \Big|_{\xi=\lambda_1\Delta} - (E[Z|\{d_Y\}])^2 \right\} \\
&= \Delta^2 \left\{ \frac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1} \left(\frac{N(N-1)e^{N\xi}}{e^{\xi}-1} - \frac{(N-1)e^{N\xi}+1}{e^{\xi}-1} - \frac{(N-1)e^{N\xi}+1}{\left(e^{\xi}-1\right)^2} \right. \right. \\
&\qquad \left. \left. - \frac{Ne^{N\xi}}{\left(e^{\xi}-1\right)^2} + 2\frac{\left(e^{N\xi}-1\right)}{\left(e^{\xi}-1\right)^2} + 2\frac{\left(e^{N\xi}-1\right)}{\left(e^{\xi}-1\right)^3} \right) \Big|_{\xi=\lambda_1\Delta} - (E[Z|\{d_Y\}])^2 \right\} \\
&= \Delta^2 \left\{ \frac{e^{\lambda_1\Delta}-1}{e^{\lambda_1\Delta N}-1} \left(\frac{(N-1)^2e^{N\xi}-1}{e^{\xi}-1} + \frac{(3-2N)e^{N\xi}-3}{\left(e^{\xi}-1\right)^2} \right. \right. \\
&\qquad \left. \left. + \frac{2\left(e^{N\xi}-1\right)}{\left(e^{\xi}-1\right)^3} \right) \Big|_{\xi=\lambda_1\Delta} - (E[Z|\{d_Y\}])^2 \right\} \\
&= \Delta^2 \left\{ \frac{(N-1)^2e^{\lambda_1\Delta N}-1}{e^{\lambda_1\Delta N}-1} + \frac{(3-2N)e^{\lambda_1\Delta N}-3}{\left(e^{\lambda_1\Delta}-1\right)\left(e^{\lambda_1\Delta N}-1\right)} + \frac{2}{\left(e^{\lambda_1\Delta}-1\right)^2} \right. \\
&\qquad \left. - \left(N-1+\frac{N}{e^{\lambda_1\Delta N}-1} - \frac{1}{e^{\lambda_1\Delta}-1} \right)^2 \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \Delta^2\Bigg\{ \tfrac{(N-1)^2\left(e^{\lambda_1\Delta N}-1+1\right)-1}{e^{\lambda_1\Delta N}-1} + \tfrac{(3-2N)\left(e^{\lambda_1\Delta N}-1+1\right)-3}{\left(e^{\lambda_1\Delta}-1\right)\left(e^{\lambda_1\Delta N}-1\right)} \\
&\qquad + \tfrac{2}{\left(e^{\lambda_1\Delta}-1\right)^2} - \Bigg((N-1)^2 + \tfrac{N^2}{\left(e^{\lambda_1\Delta N}-1\right)^2} + \tfrac{1}{\left(e^{\lambda_1\Delta}-1\right)^2} \\
&\qquad + \tfrac{2N(N-1)}{e^{\lambda_1\Delta N}-1} - \tfrac{2(N-1)}{e^{\lambda_1\Delta}-1} - \tfrac{2N}{\left(e^{\lambda_1\Delta}-1\right)\left(e^{\lambda_1\Delta N}-1\right)} \Bigg) \Bigg\} \\
&= \Delta^2\Bigg\{ (N-1)^2 + \tfrac{(N-1)^2}{e^{\lambda_1\Delta N}-1} - \tfrac{1}{e^{\lambda_1\Delta N}-1} + \tfrac{3-2N}{e^{\lambda_1\Delta}-1} \\
&\qquad - \tfrac{2N}{\left(e^{\lambda_1\Delta N}-1\right)\left(e^{\lambda_1\Delta}-1\right)} + \tfrac{2}{\left(e^{\lambda_1\Delta}-1\right)^2} - \Bigg((N-1)^2 + \tfrac{N^2}{\left(e^{\lambda_1\Delta N}-1\right)^2} \\
&\qquad + \tfrac{1}{\left(e^{\lambda_1\Delta}-1\right)^2} + \tfrac{2N(N-1)}{e^{\lambda_1\Delta N}-1} - \tfrac{2(N-1)}{e^{\lambda_1\Delta}-1} - \tfrac{2N}{\left(e^{\lambda_1\Delta}-1\right)\left(e^{\lambda_1\Delta N}-1\right)} \Bigg) \Bigg\} \\
&= \Delta^2\Bigg\{ \tfrac{(N-1)^2-2N(N-1)-1}{e^{\lambda_1\Delta N}-1} + \tfrac{3-2N+2(N-1)}{e^{\lambda_1\Delta}-1} \\
&\qquad + \tfrac{1}{\left(e^{\lambda_1\Delta}-1\right)^2} - \tfrac{N^2}{\left(e^{\lambda_1\Delta N}-1\right)^2} \Bigg\} \\
&= \Delta^2\left\{ \tfrac{1}{e^{\lambda_1\Delta}-1} + \tfrac{1}{\left(e^{\lambda_1\Delta}-1\right)^2} - \tfrac{N^2}{e^{\lambda_1\Delta N}-1} - \tfrac{N^2}{\left(e^{\lambda_1\Delta N}-1\right)^2} \right\} \\
&= \Delta^2\left\{ \tfrac{e^{\lambda_1\Delta}}{\left(e^{\lambda_1\Delta}-1\right)^2} - \tfrac{N^2\left(e^{\lambda_1\Delta N}-1\right)+N^2}{\left(e^{\lambda_1\Delta N}-1\right)^2} \right\} \\
&= \Delta^2\left\{ \tfrac{e^{\lambda_1\Delta}}{\left(e^{\lambda_1\Delta}-1\right)^2} - \tfrac{N^2 e^{\lambda_1\Delta N}}{\left(e^{\lambda_1\Delta N}-1\right)^2} \right\} \\
&= \tfrac{\Delta^2}{\left(e^{\lambda_1\Delta}-1\right)^2\left(e^{\lambda_1\Delta N}-1\right)^2}\left\{ e^{\lambda_1\Delta}\left(e^{\lambda_1\Delta N}-1\right)^2 - N^2 e^{\lambda_1\Delta N}\left(e^{\lambda_1\Delta}-1\right)^2 \right\}
\end{aligned}
\tag{7.40}
$$

and this completes the deduction proposed by the **proposition 7.1.3**. □

Corollary 7.1.6 (Standardized moments).

$$
E^s[(Y|\{d_Y\})^i] = \mu_i^{(s)} = E\left[\left(\frac{Y|\{d_Y\} - E[Y|\{d_Y\}]}{\sigma}\right)^i\right] \tag{7.41}
$$

$$
= \left(\frac{\Delta}{\sigma}\right)^i E^c\left[(Z|\{d_Y\})^i\right] \tag{7.42}
$$

$$
= \left\{(e^{\lambda_1\Delta}-1)(e^{\lambda_1\Delta N}-1)A\right\}^i E^c\left[(Z|\{d_Y\})^i\right] \tag{7.43}
$$

such that

$$A = \frac{\pm 1}{\sqrt{e^{\lambda_1\Delta}(e^{\lambda_1\Delta N}-1)^2 - N^2 e^{\lambda_1\Delta N}(e^{\lambda_1\Delta}-1)^2}} \tag{7.44}$$

where $A \gtrless 0$ *according as* $\Delta \gtrless 0$.

As special cases for $i \in \{3,4\}$, the following corollaries for **skewness** and **kurtosis** are given **without proofs**:

Corollary 7.1.7 (Skewness (the third standardized moment)).

$$\begin{aligned} E^s\left[(Y|\{d_Y\})^3\right] &= \mu_3^{(s)} = A^3\big\{e^{\lambda_1\Delta}\big(1+e^{\lambda_1\Delta}\big)\big(1-e^{3\lambda_1\Delta N}\big) \\ &\quad -N^3 e^{\lambda_1\Delta N}\big(1+e^{\lambda_1\Delta N}\big)\big(1-e^{3\lambda_1\Delta}\big) \\ &\quad +3\big(N^3-1\big)e^{\lambda_1\Delta(N+1)}\big(1-e^{\lambda_1\Delta(N+1)}\big) \\ &\quad +3\big(N^3+1\big)e^{\lambda_1\Delta(N+1)}\big(e^{\lambda_1\Delta N}-e^{\lambda_1\Delta}\big)\big\} \end{aligned} \tag{7.45}$$

Corollary 7.1.8 (Kurtosis (the fourth standardized moment)).

$$\begin{aligned} E^s\left[(Y|\{d_Y\})^4\right] &= \mu_4^{(s)} = A^4\big\{e^{\lambda_1\Delta}\big(1+e^{4\lambda_1\Delta N}\big)\big(1+7e^{\lambda_1\Delta}+e^{2\lambda_1\Delta}\big) \\ &\quad -N^4 e^{\lambda_1\Delta N}\big(1+e^{4\lambda_1\Delta}\big)\big(1+e^{\lambda_1\Delta N}+e^{2\lambda_1\Delta N}\big) \\ &\quad -6\big(N^4+4N^2-7\big)e^{2\lambda_1\Delta(N+1)} \\ &\quad +2\big(2N^4+6N^2+3\big)e^{\lambda_1\Delta(N+1)}\big(1+e^{2\lambda_1\Delta}\big) \\ &\quad -2\big(3N^4-6N^2+14\big)e^{\lambda_1\Delta(N+2)}\big(1-e^{2\lambda_1\Delta N}\big) \\ &\quad +\big(4N^4-6N^2-4\big)e^{\lambda_1\Delta(N+1)}\big(1+e^{2\lambda_1\Delta}\big)\big(1+e^{2\lambda_1\Delta N}\big)\big\} \end{aligned} \tag{7.46}$$

Corollary 7.1.9 (Median).

$$\begin{aligned} Me[Y] &= \frac{Q^{(u)}_{Y|\{d_Y\}}(0.5) + Q^{(l)}_{Y|\{d_Y\}}(0.5)}{2} \\ &= y_1 + \frac{(u-1)\Delta + (l-1)\Delta}{2} = y_1 + \left(\frac{u+l}{2} - 1\right)\Delta \end{aligned} \tag{7.47}$$

such that

$$\frac{1}{1-e^{\lambda_1\Delta N}} \leq \left(\frac{1}{1-e^{\lambda_1\Delta N}} - 0.5\right)e^{-\lambda_1\Delta u} < \frac{e^{-\lambda_1\Delta}}{1-e^{\lambda_1\Delta N}} \tag{7.48}$$

$$\frac{e^{\lambda_1\Delta}}{1-e^{\lambda_1\Delta N}} < \left(0.5 + \frac{e^{\lambda_1\Delta N}}{1-e^{\lambda_1\Delta N}}\right)e^{-\lambda_1\Delta l} \leq \frac{1}{1-e^{\lambda_1\Delta N}} \tag{7.49}$$

Proposition 7.1.4 (Stochastic Entropy).

$$H(P_{Y|\{d_Y\}}) = \log\left(\frac{1-e^{\lambda_1\Delta N}}{1-e^{\lambda_1\Delta}}\right) - \lambda_1\Delta\left(N-1+\frac{N}{e^{\lambda_1\Delta N}-1}-\frac{1}{e^{\lambda_1\Delta}-1}\right) \tag{7.50}$$

Proof of the ***proposition 7.1.4****.*

$$\begin{aligned}
H(P_{Y|\{d_Y\}}) &= -(\lambda_0+\lambda_1 E[Y|\{d_Y\}]) \\
&= -\log\left(\frac{1}{\sum_{k=1}^{N} e^{\lambda_1 y_k}}\right) \\
&\quad -\lambda_1\left\{y_1+\Delta\left(N-1+\frac{N}{e^{\lambda_1\Delta N}-1}-\frac{1}{e^{\lambda_1\Delta}-1}\right)\right\} \\
&= \log\left(e^{\lambda_1 y_1}\frac{1-e^{\lambda_1\Delta N}}{1-e^{\lambda_1\Delta}}\right) \\
&\quad -\lambda_1 y_1-\lambda_1\Delta\left(N-1+\frac{N}{e^{\lambda_1\Delta N}-1}-\frac{1}{e^{\lambda_1\Delta}-1}\right) \\
&= \log\left(\frac{1-e^{\lambda_1\Delta N}}{1-e^{\lambda_1\Delta}}\right) \\
&\quad -\lambda_1\Delta\left(N-1+\frac{N}{e^{\lambda_1\Delta N}-1}-\frac{1}{e^{\lambda_1\Delta}-1}\right)
\end{aligned} \tag{7.51}$$

and this completes the deduction proposed by the **proposition 7.1.4**. □

7.1.2 Continuous monotone probability distribution

A considerable simplification is obtained, if the usage of the discrete monotone probability distribution is replaced by the usage of the continuous monotone probability distribution. **Most** of the probability functions of the continuous monotonic probability distribution are rather elementary and therefore are given simply in form of **definitions** and their associated **corollaries**.

Definition 7.1.10 (Probability density function for $a \leq y \leq b$).

$$f_{Y|\{d_Y\}}(y) = \frac{\lambda_1 e^{\lambda_1 y}}{e^{\lambda_1 b} - e^{\lambda_1 a}} \tag{7.52}$$

Definition 7.1.11 (Distribution and survival functions for $a \leq y \leq b$).

$$F_{Y|\{d_Y\}}(y) = \frac{e^{\lambda_1 y} - e^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}} \tag{7.53}$$

$$\overline{F}_{Y|\{d_Y\}}(y) = \frac{e^{\lambda_1 b} - e^{\lambda_1 y}}{e^{\lambda_1 b} - e^{\lambda_1 a}} \tag{7.54}$$

Definition 7.1.12 (Quantile functions for $0 < z \leq 1$).

$$Q^{(u)}_{Y|\{d_Y\}}(z) = y_z \tag{7.55}$$

with

$$y_z = a + \frac{1}{\lambda_1} \log\left[1 + z\left(e^{\lambda_1 (b-a)} - 1\right)\right] \tag{7.56}$$

simply because $$F_{Y|\{d_Y\}}(y_z) = \frac{e^{\lambda_1 y_z} - e^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}} = z$$

and

$$Q^{(\ell)}_{Y|\{d_Y\}}(z) = y_z \tag{7.57}$$

with

$$y_z = b + \frac{1}{\lambda_1} \log\left[1 - z\left(1 - e^{-\lambda_1 (b-a)}\right)\right] \tag{7.58}$$

simply because $$\overline{F}_{Y|\{d_Y\}}(y_z) = \frac{e^{\lambda_1 b} - e^{\lambda_1 y_z}}{e^{\lambda_1 b} - e^{\lambda_1 a}} = z$$

Definition 7.1.13 (Failure intensity function).

$$a_{Y|\{d_Y\}}(y) = \frac{f_{Y|\{d_Y\}}(y)}{\overline{F}_{Y|\{d_Y\}}(y)} = \frac{\lambda_1}{e^{\lambda_1(b-y)} - 1} \tag{7.59}$$

Definition 7.1.14 (Moments).

$$E[(Y|\{d_Y\})^i] = \int_a^b y^i \frac{\lambda_1 e^{\lambda_1 y}}{e^{\lambda_1 b} - e^{\lambda_1 a}} dy \tag{7.60}$$

Corollary 7.1.10 (Recursive relation between moments). *From (7.60), the general recursive relation between the moments can be deduced (for every $i \in \mathbb{N}$) as*

$$E[(Y|\{d_Y\})^i] = \frac{b^i e^{\lambda_1 b} - a^i e^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}} - \frac{i}{\lambda_1} E[(Y|\{d_Y\})^{i-1}] \tag{7.61}$$

As a special case of (7.60) for $i = 1$,

Corollary 7.1.11 (The first moment).

$$E[Y|\{d_Y\}] = \frac{be^{\lambda_1 b} - ae^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}} - \frac{1}{\lambda_1} \tag{7.62}$$

Definition 7.1.15 (Central moments).

$$\begin{aligned} E^c[(Y|\{d_Y\})^i] &= E[(Y|\{d_Y\} - E[Y|\{d_Y\}])^i] && (7.63) \\ &= \int_a^b (y-\mu)^i \frac{\lambda_1 e^{\lambda_1 y}}{e^{\lambda_1 b} - e^{\lambda_1 a}} dy && (7.64) \end{aligned}$$

Corollary 7.1.12 (Recursive relation between central moments). *From (7.64), the general recursive relation between the central moments can be deduced (for every $i \in \mathbb{N}$) as*

$$\begin{aligned} E^c[(Y|\{d_Y\})^i] &= \frac{(b-\mu)^i e^{\lambda_1 b} - (a-\mu)^i e^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}} - \frac{i}{\lambda} E^c[(Y|\{d_Y\})^{i-1}] \\ &= \frac{\left(\frac{1}{\lambda} - \frac{(b-a)e^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}}\right)^i e^{\lambda_1 b} - \left(\frac{1}{\lambda} - \frac{(b-a)e^{\lambda_1 b}}{e^{\lambda_1 b} - e^{\lambda_1 a}}\right)^i e^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}} \\ &\quad - \frac{i}{\lambda} E^c[(Y|\{d_Y\})^{i-1}] \end{aligned} \tag{7.65}$$

As a special case of (7.60) for $i = 2$, (of course $E^c[Y|\{d_Y\}] = 0$)

Corollary 7.1.13 (Variance (the second central moment)).

$$\begin{aligned} V[Y|\{d_Y\}] &= E^c[(Y|\{d_Y\})^2] \\ &= \frac{\left(\frac{1}{\lambda} - \frac{(b-a)e^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}}\right)^2 e^{\lambda_1 b} - \left(\frac{1}{\lambda} - \frac{(b-a)e^{\lambda_1 b}}{e^{\lambda_1 b} - e^{\lambda_1 a}}\right)^2 e^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}} \\ &= \frac{\left\{(e^{\lambda_1 b} - e^{\lambda_1 a}) - \lambda_1(b-a)e^{\lambda_1 a}\right\}^2 e^{\lambda_1 b}}{\lambda_1^2(e^{\lambda_1 b} - e^{\lambda_1 a})^3} \\ &\quad - \frac{\left\{(e^{\lambda_1 b} - e^{\lambda_1 a}) - \lambda_1(b-a)e^{\lambda_1 b}\right\}^2 e^{\lambda_1 a}}{\lambda_1^2(e^{\lambda_1 b} - e^{\lambda_1 a})^3} \\ &= \frac{1}{\lambda_1^2} - \frac{(b-a)^2 e^{\lambda_1(a+b)}}{(e^{\lambda_1 b} - e^{\lambda_1 a})^2} \end{aligned} \tag{7.66}$$

Therefore, the standard deviation (σ_Y) of $Y|\{d_Y\}$ is given by

Corollary 7.1.14 (The standard deviation).

$$\sigma_Y = \frac{\sqrt{(e^{\lambda_1 b} - e^{\lambda_1 a})^2 - \lambda_1^2 (b-a)^2 e^{\lambda_1(a+b)}}}{\lambda_1 (e^{\lambda_1 b} - e^{\lambda_1 a})} = \frac{\Delta}{\lambda_1 (e^{\lambda_1 b} - e^{\lambda_1 a})} \tag{7.67}$$

by putting

$$\Delta = \sqrt{(e^{\lambda_1 b} - e^{\lambda_1 a})^2 - \lambda_1^2 (b-a)^2 e^{\lambda_1(a+b)}} \tag{7.68}$$

Definition 7.1.16 (Standardized moments).

$$E^s[(Y|\{d_Y\})^i] = \int_a^b \left(\frac{y - \mu_Y^{(1)}}{\sigma_Y}\right)^i \frac{\lambda_1 e^{\lambda_1 y}}{e^{\lambda_1 b} - e^{\lambda_1 a}} dy \tag{7.69}$$

Corollary 7.1.15 (Recursive relation between standardized moments). *From (7.69), (7.67) and (7.68) the general recursive relation between the standardized moments can be deduced (for every $i \in \mathbb{N}$) as*

$$\begin{aligned} E^s[(Y|\{d_Y\})^i] &= \frac{\left(\frac{e^{\lambda_1 b} - e^{\lambda_1 a} - \lambda_1(b-a)e^{\lambda_1 a}}{\Delta}\right)^i e^{\lambda_1 b}}{e^{\lambda_1 b} - e^{\lambda_1 a}} \\ &\quad - \frac{\left(\frac{e^{\lambda_1 b} - e^{\lambda_1 a} - \lambda_1(b-a)e^{\lambda_1 b}}{\Delta}\right)^i e^{\lambda_1 a}}{e^{\lambda_1 b} - e^{\lambda_1 a}} \\ &\quad - \frac{i\left(e^{\lambda_1 b} - e^{\lambda_1 a}\right)}{\Delta} E^s[(Y|\{d_Y\})^{i-1}] \end{aligned} \tag{7.70}$$

As a special cases of (7.60) for $i \in \{3,4\}$, (of course $E^s[Y|\{d_Y\}] = 0$)

Proposition 7.1.5 (Skewness (the third standardized moment)).

$$Sk[Y|\{d_Y\}] = \frac{1}{\Delta^3}\left\{\lambda_1^3(b-a)^3 e^{\lambda_1(a+b)}\left(e^{\lambda_1 b}+e^{\lambda_1 a}\right) - 2\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^3\right\} \tag{7.71}$$

Proof of the ***proposition 7.1.5****.*

$$\begin{aligned}
Sk[Y|\{d_Y\}] &= E^s\left[(Y|\{d_Y\})^3\right] \\
&= \frac{\left(\frac{e^{\lambda_1 b}-e^{\lambda_1 a}-\lambda_1(b-a)e^{\lambda_1 a}}{\Delta}\right)^3 e^{\lambda_1 b} - \left(\frac{e^{\lambda_1 b}-e^{\lambda_1 a}-\lambda_1(b-a)e^{\lambda_1 b}}{\Delta}\right)^3 e^{\lambda_1 a}}{e^{\lambda_1 b}-e^{\lambda_1 a}} \\
&\quad -\frac{3}{\Delta}(e^{\lambda_1 b}-e^{\lambda_1 a}) \\
&= \frac{1}{\Delta^3(e^{\lambda_1 b}-e^{\lambda_1 a})}\left[(e^{\lambda_1 b}-e^{\lambda_1 a})^3(e^{\lambda_1 b}-e^{\lambda_1 a})\right. \\
&\quad +3(e^{\lambda_1 b}-e^{\lambda_1 a})\left\{\lambda_1^2(b-a)^2 e^{2\lambda_1 a+\lambda_1 b} - \lambda_1^2(b-a)^2 e^{2\lambda_1 b+\lambda_1 a}\right\} \\
&\quad \left.-\left\{\lambda_1^3(b-a)^3 e^{3\lambda_1 a+\lambda_1 b} - \lambda_1^3(b-a)^3 e^{3\lambda_1 b+\lambda_1 a}\right\}\right] \\
&\quad -\frac{3}{\Delta}(e^{\lambda_1 b}-e^{\lambda_1 a}) \\
&= \frac{(e^{\lambda_1 b}-e^{\lambda_1 a})^3 - 3(e^{\lambda_1 b}-e^{\lambda_1 a})\left\{\lambda_1^2(b-a)^2 e^{\lambda_1(a+b)}\right\}}{\Delta^3} \\
&\quad +\frac{\lambda_1^3(b-a)^3 e^{\lambda_1(a+b)}\left(e^{\lambda_1 b}+e^{\lambda_1 a}\right)}{\Delta^3} - \frac{3}{\Delta}(e^{\lambda_1 b}-e^{\lambda_1 a}) \\
&= \frac{1}{\Delta^3}\left[(e^{\lambda_1 b}-e^{\lambda_1 a})^3 - 3(e^{\lambda_1 b}-e^{\lambda_1 a})\left\{\lambda_1^2(b-a)^2 e^{\lambda_1(a+b)}\right\}\right. \\
&\quad +\lambda_1^3(b-a)^3 e^{\lambda_1(a+b)}\left(e^{\lambda_1 b}+e^{\lambda_1 a}\right) \\
&\quad \left.-3\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)\left\{\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^2 - \lambda_1^2(b-a)^2 e^{\lambda_1(a+b)}\right\}\right] \\
&= \frac{1}{\Delta^3}\left\{\lambda_1^3(b-a)^3 e^{\lambda_1(a+b)}\left(e^{\lambda_1 b}+e^{\lambda_1 a}\right) - 2\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^3\right\}
\end{aligned}$$

and this completes the deduction proposed by the **proposition 7.1.5**. □

Proposition 7.1.6 (Kurtosis (the fourth standardized moment)).

$$Ku[Y|\{d_Y\}] = \frac{1}{\Delta^4}\Big\{9\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^4 - 6\lambda_1^2(b-a)^2 e^{\lambda_1(a+b)}\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^2 \\ -\lambda_1^4(b-a)^4 e^{\lambda_1(a+b)}\left[e^{2\lambda_1 b}+e^{\lambda_1(a+b)}+e^{2\lambda_1 a}\right]\Big\} \tag{7.72}$$

Proof of the ***proposition 7.1.6****.*

$$\begin{aligned} Ku[Y|\{d_Y\}] &= E^s\left[(Y|\{d_Y\})^4\right] \\ &= \frac{\left(\frac{e^{\lambda_1 b}-e^{\lambda_1 a}-\lambda_1(b-a)e^{\lambda_1 a}}{\Delta}\right)^4 e^{\lambda_1 b} - \left(\frac{e^{\lambda_1 b}-e^{\lambda_1 a}-\lambda_1(b-a)e^{\lambda_1 b}}{\Delta}\right)^4 e^{\lambda_1 a}}{e^{\lambda_1 b}-e^{\lambda_1 a}} \\ &\quad -\frac{4}{\Delta}\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)E^s[Y^3] \\ &= \frac{1}{\Delta^4\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)}\Big\{\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^4\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right) \\ &\quad +6\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^2\Big[\left\{\lambda_1(b-a)e^{\lambda_1 a}\right\}^2 e^{\lambda_1 b} \\ &\qquad -\left\{\lambda_1(b-a)e^{\lambda_1 b}\right\}^2 e^{\lambda_1 a}\Big] \\ &\quad -4\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)\Big[\left\{\lambda_1(b-a)e^{\lambda_1 a}\right\}^3 e^{\lambda_1 b} \\ &\qquad -\left\{\lambda_1(b-a)e^{\lambda_1 b}\right\}^3 e^{\lambda_1 a}\Big] \\ &\quad +\Big[\left\{\lambda_1(b-a)e^{\lambda_1 a}\right\}^4 e^{\lambda_1 b} - \left\{\lambda_1(b-a)e^{\lambda_1 b}\right\}^4 e^{\lambda_1 a}\Big]\Big\} \\ &\quad -\frac{4}{\Delta}\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)\frac{1}{\Delta^3}\Big\{\lambda_1^3(b-a)^3 e^{\lambda_1(a+b)}\left(e^{\lambda_1 b}+e^{\lambda_1 a}\right) \\ &\qquad -2\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^3\Big\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\Delta^4}\left\{\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^4-6\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^2\lambda_1^2(b-a)^2e^{\lambda_1(a+b)}\right.\\
&\quad +4\left[\lambda_1^3(b-a)^3e^{\lambda_1(a+b)}\left\{e^{2\lambda_1 b}-e^{2\lambda_1 a}\right\}\right]\\
&\quad \left.-\lambda_1^4(b-a)^4e^{\lambda_1(a+b)}\left[e^{2\lambda_1 b}+e^{\lambda_1(a+b)}+e^{2\lambda_1 a}\right]\right\}\\
&\quad -\frac{4\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)}{\Delta^4}\left[\lambda_1^3(b-a)^3e^{\lambda_1(a+b)}\left(e^{\lambda_1 b}+e^{\lambda_1 a}\right)\right.\\
&\qquad \left.-2\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^3\right]\\
&= \frac{1}{\Delta^4}\left\{9\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^4-6\lambda_1^2(b-a)^2e^{\lambda_1(a+b)}\left(e^{\lambda_1 b}-e^{\lambda_1 a}\right)^2\right.\\
&\quad \left.-\lambda_1^4(b-a)^4e^{\lambda_1(a+b)}\left[e^{2\lambda_1 b}+e^{\lambda_1(a+b)}+e^{2\lambda_1 a}\right]\right\}
\end{aligned}
$$

and this completes the deduction proposed by the **proposition 7.1.6**. □

Definition 7.1.17 (Median).

$$
\begin{aligned}
Me[Y] &= \frac{Q_Y^{(u)}(0.5)+Q_Y^{(l)}(0.5)}{2}\\
&= \frac{a+\frac{1}{\lambda_1}\log\left[1+\frac{1}{2}\left(e^{\lambda_1(b-a)}-1\right)\right]}{2}\\
&\qquad +\frac{b+\frac{1}{\lambda_1}\log\left[1-\frac{1}{2}\left(1-e^{-\lambda_1(b-a)}\right)\right]}{2}\\
&= \frac{a+b}{2}+\frac{1}{\lambda_1}\log\left[\cosh\left(\frac{\lambda_1}{2}(b-a)\right)\right]
\end{aligned}
$$

Definition 7.1.18 (Stochastic entropy). *By using (7.62), the stochastic entropy can be easily defined as follows:*

$$
\begin{aligned}
H\left(f_{Y|\{d_Y\}}\right) &= -\lambda_0-\lambda_1 E[Y|\{d_Y\}]\\
&= 1+\log\left(\frac{e^{\lambda_1 b}-e^{\lambda_1 a}}{\lambda_1}\right)-\lambda_1\left(\frac{be^{\lambda_1 b}-ae^{\lambda_1 a}}{e^{\lambda_1 b}-e^{\lambda_1 a}}\right)
\end{aligned}
\tag{7.73}
$$

7.2 Uni-extremal probability distribution

7.2.1 Discrete uni-extremal probability distribution

As usual, we assume that the range of variability of $Y|\{d_Y\}$

$$\mathcal{X}_Y(\{d_Y\}) = \{y_1, y_2, \ldots, y_N\} \tag{7.74}$$

is known, the elements of which are arranged in the ascending order, i.e., $y_1 < y_2 < \ldots < y_N$. In this case, the minimum information probability distribution is, in general, an approximation. For obtaining it, besides the range of variability, the actual value $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$ (i.e. the numerical values of the first and the second moment) are needed. The minimum information distribution is given by the following probability mass function of $Y|\{d_Y\}$:

$$f_{Y|\{d_Y\}}(y) = e^{\lambda_0 + \lambda_1 y + \lambda_2 y^2} \text{ for } y \in \mathcal{X}_Y(\{d_Y\}) \tag{7.75}$$

with λ_0, λ_1 and $\lambda_2 \neq 0$ uniquely determined by the solution of the following three equations:

$$\sum_{j=1}^{N} e^{\lambda_0 + \lambda_1 y_j + \lambda_2 y_j^2} = 1 \tag{7.76}$$

$$\sum_{j=1}^{N} y_j e^{\lambda_0 + \lambda_1 y_j + \lambda_2 y_j^2} = \mu_Y^{(1)} \tag{7.77}$$

$$\sum_{j=1}^{N} y_j^2 e^{\lambda_0 + \lambda_1 y_j + \lambda_2 y_j^2} = \mu_Y^{(2)} \tag{7.78}$$

Here,

- $\lambda_2 > 0$ is the **necessary condition** for the probability distribution to be a bathtub-shaped probability mass function.
- $\lambda_2 < 0$ is the **necessary condition** for the probability distribution to be a uni-modal probability mass function.

Let us put $M = -\frac{\lambda_1}{2\lambda_2}$. Then, it is evidently clear that $f_{Y|\{d_Y\}}(M)$ could be the global extremum value of the probability mass function, **provided** $M \in \mathcal{X}_Y(\{d_Y\})$.

The probability distribution (7.75) has the following characteristic properties:

Definition 7.2.1 (Probability mass function for $j = 1, 2, \ldots, N$).

$$f_{Y|\{d_Y\}}(y_j) = \frac{e^{\lambda_2(y_j - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2(y_k - M)^2}} \tag{7.79}$$

Definition 7.2.2 (Distribution and survival functions).

$$F_{Y|\{d_Y\}}(y_j) = \frac{\sum_{k=1}^{j} e^{\lambda_2(y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2(y_k - M)^2}} \tag{7.80}$$

$$\overline{F}_{Y|\{d_Y\}}(y_j) = \frac{\sum_{k=j}^{N} e^{\lambda_2(y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2(y_k - M)^2}} \tag{7.81}$$

Definition 7.2.3 (Quantile functions). *For $z \in (0, 1]$, the upper and the lower quantile functions are given as:*

$$Q^{(u)}_{Y|\{d_Y\}}(z) = y_{j_1} \tag{7.82}$$

such that j_1 is given by

- $\frac{\sum_{k=1}^{j_1 - 1} e^{\lambda_2(y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2(y_k - M)^2}} < z \leq \frac{\sum_{k=1}^{j_1} e^{\lambda_2(y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2(y_k - M)^2}}$

and

$$Q^{(\ell)}_{Y|\{d_Y\}}(z) = y_{j_2} \tag{7.83}$$

such that j_2 is given by

- $\frac{\sum_{k=j_2+1}^{N} e^{\lambda_2(y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2(y_k - M)^2}} < z \leq \frac{\sum_{k=j_2}^{N} e^{\lambda_2(y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2(y_k - M)^2}}$

Definition 7.2.4 (Failure intensity function).

$$a_{Y|\{d_Y\}}(y_j) = \frac{f_{Y|\{d_Y\}}(y_j)}{\overline{F}_{Y|\{d_Y\}}(y_j)} = \frac{e^{\lambda_2(y_j - M)^2}}{\sum_{k=j}^{N} e^{\lambda_2(y_k - M)^2}} \tag{7.84}$$

Definition 7.2.5 (Moments).

$$E[(Y|\{d_Y\})^i] = \frac{\sum_{k=1}^{N} y_k^i e^{\lambda_2 (y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2 (y_k - M)^2}} \tag{7.85}$$

Definition 7.2.6 (Central moments).

$$\begin{aligned} E^c[(Y|\{d_Y\})^i] &= E\left[\Big(Y|\{d_Y\} - E[Y|\{d_Y\}]\Big)^i\right] \\ &= \frac{\sum_{k=1}^{N} \Big(y_k - E[Y|\{d_Y\}]\Big)^i e^{\lambda_2 (y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2 (y_k - M)^2}} \end{aligned} \tag{7.86}$$

Definition 7.2.7 (Standardized moments).

$$\begin{aligned} E^s[(Y|\{d_Y\})^i] &= E\left[\left(\frac{Y|\{d_Y\} - E[Y|\{d_Y\}]}{\sigma}\right)^i\right] \\ &= \frac{\sum_{k=1}^{N} \left(\frac{y_k - E[Y|\{d_Y\}]}{\sigma_Y}\right)^i e^{\lambda_2 (y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2 (y_k - M)^2}} \end{aligned} \tag{7.87}$$

such that $\sigma_Y = \sqrt{E^c[(Y|\{d_Y\})^2]}$

Definition 7.2.8 (Median).

$$Me[Y] = \frac{Q_Y^{(u)}(0.5) + Q_Y^{(\ell)}(0.5)}{2} \tag{7.88}$$

Definition 7.2.9 (Stochastic entropy).

$$\begin{aligned} H\left(P_{Y|\{d_Y\}}\right) &= -\left(\lambda_0 + \lambda_1 E[Y|\{d_Y\}] + \lambda_2 E[(Y|\{d_Y\})^2]\right) \\ &= -\left(\lambda_0 + \lambda_1 \frac{\sum_{k=1}^{N} y_k e^{\lambda_2 (y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2 (y_k - M)^2}} + \lambda_2 \frac{\sum_{k=1}^{N} y_k^2 e^{\lambda_2 (y_k - M)^2}}{\sum_{k=1}^{N} e^{\lambda_2 (y_k - M)^2}}\right) \end{aligned} \tag{7.89}$$

If the elements of the range of variability of $Y|\{d_Y\}$ follow an arithmetic progression, the probability functions and the moments of special interest do not yield any spectacular simplified results in this case.

Therefore, the **same formulae** as above should be used for that.

7.2.2 Continuous uni-extremal probability distribution

Unlike **continuous monotone** probability distribution, a **continuous uni-extremal** probability distribution **demands numerical integrations** for the calculation of probability functions and moments of $Y|\{d_Y\}$.

For the sake of convenience of calculating the oncoming expressions, we make use of $\lambda_1 = \frac{\beta}{b-a} - \frac{2a\gamma}{(b-a)^2}$ and $\lambda_2 = \frac{\gamma}{(b-a)^2}$. Obviously $\lambda_2 \neq 0 \Leftrightarrow \gamma \neq 0$, which shall be our basic assumption throughout this subsection. So, **with subject to the usage of** β, γ, a and b, we proceed to **define** the parameters of the continuous uni-extremal probability distribution as follows:

Definition 7.2.10 (Probability density function for $a \leq y \leq b$).

$$f_{Y|\{d_Y\}} = \frac{e^{\beta\left(\frac{y-a}{b-a}\right)+\gamma\left(\frac{y-a}{b-a}\right)^2}}{(b-a)\int\limits_0^1 e^{\beta t+\gamma t^2}dt}, \quad a < y < b \tag{7.90}$$

It should be noted, that

$$\begin{aligned} e^{\lambda_0} &= \frac{e^{\gamma\left(\frac{a}{b-a}\right)^2-\beta\left(\frac{a}{b-a}\right)}}{(b-a)\int\limits_0^1 e^{\beta t+\gamma t^2}dt} \\ \Rightarrow \lambda_0 &= \gamma\left(\frac{a}{b-a}\right)^2 - \beta\left(\frac{a}{b-a}\right) - \log\left((b-a)\int\limits_0^1 e^{\beta t+\gamma t^2}dt\right) \end{aligned} \tag{7.91}$$

Definition 7.2.11 (Distribution and survival functions for $a \leq y \leq b$).

$$F_{Y|\{d_Y\}}(y) = \frac{\int\limits_0^{\frac{y-a}{b-a}} e^{\beta t+\gamma t^2}dt}{\int\limits_0^1 e^{\beta t+\gamma t^2}dt} \tag{7.92}$$

$$\overline{F}_{Y|\{d_Y\}}(y) = \frac{\int\limits_{\frac{y-a}{b-a}}^1 e^{\beta t+\gamma t^2}dt}{\int\limits_0^1 e^{\beta t+\gamma t^2}dt} \tag{7.93}$$

Definition 7.2.12 (Quantile functions for $0 < z \leq 1$).

$$Q^{(u)}_{Y|\{d_Y\}}(z) = y_z \tag{7.94}$$

for which

$$\int_0^{\frac{y_z-a}{b-a}} e^{\beta t+\gamma t^2} dt = z \int_0^1 e^{\beta t+\gamma t^2} dt \tag{7.95}$$

and

$$Q^{(\ell)}_{Y|\{d_Y\}}(z) = y_z \tag{7.96}$$

for which

$$\int_{\frac{y_z-a}{b-a}}^1 e^{\beta t+\gamma t^2} dt = z \int_0^1 e^{\beta t+\gamma t^2} dt \tag{7.97}$$

Remark 7.2.1 (On quantiles). *We can very well see, that the computation of upper and lower quantiles require a numerical solution of the above integral equations (7.95) and (7.97) (equations in y_z for given values of z), where numerical treatment on integrations are necessary.*

Definition 7.2.13 (Failure intensity function).

$$a_{Y|\{d_Y\}}(y) = \frac{f_{Y|\{d_Y\}}(y)}{\overline{F}_{Y|\{d_Y\}}(y)} = \frac{e^{\beta\left(\frac{y-a}{b-a}\right)+\gamma\left(\frac{y-a}{b-a}\right)^2}}{(b-a)\int\limits_{\frac{y-a}{b-a}}^1 e^{\beta t+\gamma t^2} dt} \tag{7.98}$$

Right at this point, we need to mention a very important thing regarding the moments of Y: For evaluation of moments of $Y|\{d_Y\}$ in this case, the introduction of the transformed random variable $X|\{d\}$ is necessary at first. So, **utilizable expressions** of the moments of $X|\{d\}$ are of **absolute necessity**.

The transformed random variable $X|\{d\} = \frac{Y|\{d_Y\}-a}{b-a}$, whose range of variability is $[0, 1]$, must have the following probability density function

$$f_{X|\{d\}} = \frac{e^{\beta x+\gamma x^2}}{\int\limits_0^1 e^{\beta t+\gamma t^2} dt} \tag{7.99}$$

With this, we proceed to give a **utilizable deduction** of the first moment of $X|\{d\}$

Proposition 7.2.1 (The first moment of X). *For $\gamma \neq 0$,*

$$\mu_1 = \frac{1}{2\gamma}\left\{\frac{e^{\beta+\gamma}-1}{\int\limits_0^1 e^{\beta x+\gamma x^2}dx} - \beta\right\} = \frac{1}{2\gamma}\left\{\frac{1-e^{-(\beta+\gamma)}}{\int\limits_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx} - \beta\right\}$$

Proof of the ***proposition 7.2.1****.*

$$\begin{aligned}
\mu_1 &= \frac{\int\limits_0^1 x e^{\beta x+\gamma x^2}dx}{\int\limits_0^1 e^{\beta x+\gamma x^2}dx} \\
&= \frac{\left(e^{\beta x}\frac{1}{2\gamma}e^{\gamma x^2}\right)\Big|_{x=0}^{x=1} - \frac{\beta}{2\gamma}\int\limits_0^1 e^{\beta x+\gamma x^2}dx}{\int\limits_0^1 e^{\beta x+\gamma x^2}dx} \\
&= \frac{\frac{1}{2\gamma}\left(e^{\beta+\gamma}-1\right) - \frac{\beta}{2\gamma}\int\limits_0^1 e^{\beta x+\gamma x^2}dx}{\int\limits_0^1 e^{\beta x+\gamma x^2}dx} \\
&= \frac{1}{2\gamma}\left\{\frac{e^{\beta+\gamma}-1}{\int\limits_0^1 e^{\beta x+\gamma x^2}dx} - \beta\right\} \qquad (7.100) \\
&= \frac{1}{2\gamma}\left\{\frac{1-e^{-(\beta+\gamma)}}{\int\limits_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx} - \beta\right\} \qquad (7.101)
\end{aligned}$$

and this completes the deduction proposed by the **proposition 7.2.1**. □

Therefore, with the help of the **first moment** of $X|\{d\}$, we proceed to give an utilizable deduction of the **second moment** of $X|\{d\}$ as follows:

Proposition 7.2.2 (The second moment of X). *For $\gamma \neq 0$,*

$$\mu_2 = \frac{1}{2\gamma}\left\{\frac{e^{\beta+\gamma}}{\int_0^1 e^{\beta x+\gamma x^2}dx} - 1 - \beta\mu_1\right\} = \frac{1}{2\gamma}\left\{\frac{1}{\int_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx} - 1 - \beta\mu_1\right\}$$

Proof of the ***proposition 7.2.2****.*

$$\begin{aligned}
\mu_2 &= \frac{\int_0^1 x^2 e^{\beta x+\gamma x^2}dx}{\int_0^1 e^{\beta x+\gamma x^2}dx} \\
&= \frac{\left(xe^{\beta x}\frac{1}{2\gamma}e^{\gamma x^2}\right)\Big|_{x=0}^{x=1} - \int_0^1 (e^{\beta x}+\beta x e^{\beta x})\frac{1}{2\gamma}e^{\gamma x^2}dx}{\int_0^1 e^{\beta x+\gamma x^2}dx} \\
&= \frac{\frac{1}{2\gamma}e^{\beta+\gamma} - \frac{1}{2\gamma}\int_0^1 e^{\beta x+\gamma x^2}dx - \frac{\beta}{2\gamma}\int_0^1 xe^{\beta x+\gamma x^2}dx}{\int_0^1 e^{\beta x+\gamma x^2}dx} \\
&= \frac{1}{2\gamma}\left\{\frac{e^{\beta+\gamma}}{\int_0^1 e^{\beta x+\gamma x^2}dx} - 1 - \beta\mu_1\right\} \qquad (7.102) \\
&= \frac{1}{2\gamma}\left\{\frac{1}{\int_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx} - 1 - \beta\mu_1\right\} \qquad (7.103)
\end{aligned}$$

and this completes the deduction proposed by the **proposition 7.2.2**. □

Therefore, on the basis of the these **first two moments** of $X|\{d\}$, the higher order moments of $X|\{d\}$ greater than 2 (i.e for $n > 2$) can be deduced **recursively** as follows:

Proposition 7.2.3 (The n^{th} moment of X, $n \in \mathbf{N}$). *For $\gamma \neq 0$,*

$$\mu_n = \frac{1}{2\gamma}\left\{\frac{e^{\beta+\gamma}}{\int_0^1 e^{\beta x+\gamma x^2}dx} - \beta\mu_{n-1} - (n-1)\mu_{n-2}\right\}$$
$$= \frac{1}{2\gamma}\left\{\frac{1}{\int_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx} - \beta\mu_{n-1} - (n-1)\mu_{n-2}\right\}$$

Proof of the ***proposition 7.2.3****.*

$$\mu_n = \frac{\int_0^1 x^n e^{\beta x+\gamma x^2}dx}{\int_0^1 e^{\beta x+\gamma x^2}dx}$$
$$= \frac{\left(x^{n-1}e^{\beta x}\frac{1}{2\gamma}e^{\gamma x^2}\right)\Big|_{x=0}^{x=1} - \int_0^1 \left(\beta x^{n-1} + (n-1)x^{n-2}\right)e^{\beta x}\frac{1}{2\gamma}e^{\gamma x^2}dx}{\int_0^1 e^{\beta x+\gamma x^2}dx}$$
$$= \frac{\frac{1}{2\gamma}e^{\beta+\gamma} - \frac{\beta}{2\gamma}\int_0^1 x^{n-1}e^{\beta x+\gamma x^2}dx - \frac{n-1}{2\gamma}\int_0^1 x^{n-2}e^{\beta x+\gamma x^2}dx}{\int_0^1 e^{\beta x+\gamma x^2}dx}$$
$$= \frac{1}{2\gamma}\left\{\frac{e^{\beta+\gamma}}{\int_0^1 e^{\beta x+\gamma x^2}dx} - \beta\mu_{n-1} - (n-1)\mu_{n-2}\right\} \tag{7.104}$$
$$= \frac{1}{2\gamma}\left\{\frac{1}{\int_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx} - \beta\mu_{n-1} - (n-1)\mu_{n-2}\right\} \tag{7.105}$$

and this completes the deduction proposed by the **proposition 7.2.3**. □

Remark 7.2.2 (Recursive computation of moments of X and handling of the overflow error problem). *With the help of (7.100) and (7.102), any higher order moment of $X|\{d\}$ (for $n > 2$) can be computed recursively by (7.104).*

*As far as our **programming work** is concerned, cases often arise when the value of the real number $e^{\beta+\gamma}$ is **too large** for the execution of the program resulting in **overflow errors**. Only in such cases, as an alternative approach, the formulae (7.101), (7.103) and consequently (7.105) should be used instead.*

Remark 7.2.3 (Skillful handling of numerical integration). *A very important thing to be noted is, that for the computation of a moment of $X|\{d\}$ **of any order**, the **numerical integration is needed exactly once**. That is, either the integral $\int_0^1 e^{\beta x+\gamma x^2}dx$ or the integral $\int_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx$ (as the case may be) needs to be computed during the computational procedure just once and **no more** numerical integrations are necessary for that.*

Remark 7.2.4 (Moments of Y). *Therefore, with the help of (7.100), (7.102) and (7.104) or alternatively with the help of (7.101), (7.103) and (7.105), the computation of a moment of $Y|\{d_Y\}$ of any order (i.e. for any $i \in \mathbb{N}$) is possible by*

$$\begin{aligned} E[(Y|\{d_Y\})^i] &= a^i + ia^{i-1}(b-a)E[X|\{d\}] \\ &\quad + \binom{i}{2}a^{i-2}(b-a)^2E[(X|\{d\})^2] \\ &\quad + \ldots + (b-a)^iE[(X|\{d\})^i] \end{aligned} \tag{7.106}$$

Remark 7.2.5 (Skillful utility of the random variable X). *For the computation of a moment of $Y|\{d_Y\}$ of **any** order described in (7.106), it has to be noted, that the **numerical computation** of **either** $\int_0^1 e^{\beta x+\gamma x^2}dx$ **or** $\int_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx$ is needed **only**.*

The introduction of the transformed random variable $X|\{d\}$ and the real numbers β and γ for the purpose of computation of the moments of $Y|\{d_Y\}$ makes sense, simply because the ***direct*** *procedure defined by*

$$E[(Y|\{d_Y\})^i] = \frac{\int_a^b y^i e^{\lambda_1 y+\lambda_2 y^2} dy}{\int_a^b e^{\lambda_1 y+\lambda_2 y^2} dy}$$

is undoubtedly more ***complicated****,* ***time consuming*** *and* ***unprotected against overflow errors****.*

For our coming discussions, we shall make use of the notations $E[X|\{d\}] = \mu_1$ and $E[(X|\{d\})^2] = \mu_2$.

Definition 7.2.14 (The first moment of Y). *For $i = 1$,*

$$E[Y|\{d_Y\}] = \mu_Y^{(1)} = a + (b-a)E[X|\{d\}] = a + (b-a)\mu_1 \tag{7.107}$$

Definition 7.2.15 (Central moments of Y). *For $i \in \mathbb{N}$,*

$$\begin{aligned} E^c[(Y|\{d_Y\})^i] &= (b-a)^i E[(X|\{d\} - \mu_1)^i] \\ &= (b-a)^i \Big\{ E[(X|\{d\})^i] - iE[(X|\{d\})^{i-1}]\mu_1 \\ &\qquad + \binom{i}{2} E[(X|\{d\})^{i-2}]\mu_1^2 + \ldots + (-1)^i \mu_1^i \Big\} \end{aligned} \tag{7.108}$$

and as a special case for $i = 2$, we have

Corollary 7.2.1 (Variance (the second central moment of Y)).

$$V[Y|\{d_Y\}] = E^c[(Y|\{d_Y\})^2] = (b-a)^2(\mu_2 - \mu_1^2) \tag{7.109}$$

and

Corollary 7.2.2 (Standard deviation of Y). *Therefore, the standard deviation (σ_Y) of $Y|\{d_Y\}$ is given by*

$$\sigma_Y = (b-a)\sqrt{\mu_2 - \mu_1^2} = (b-a)\sigma \tag{7.110}$$

Definition 7.2.16 (Standardized moments). *We know, that the standardized moments of Y and X are the* ***same****. So, for $i \in \mathbb{N}$,*

$$\begin{aligned} E^s[(Y|\{d_Y\})^i] &= \left(\frac{1}{\sigma}\right)^i E[(X|\{d\} - \mu_1)^i] \\ &= \left(\frac{1}{\sigma}\right)^i \Big\{ E[(X|\{d\})^i] - iE[(X|\{d\})^{i-1}]\mu_1 \\ &\qquad + \binom{i}{2} E[(X|\{d\})^{i-2}]\mu_1^2 + \ldots + (-1)^i \mu_1^i \Big\} \end{aligned} \tag{7.111}$$

By taking $\mu_3 = E[(X|\{d\})^3]$ and $\mu_4 = E[(X|\{d\})^4]$ as special cases, for $i \in \{3, 4\}$, we have

Corollary 7.2.3 (Skewness (the third standardized moment)).

$$\begin{aligned} Sk[Y|\{d_Y\}] &= E^s[(Y|\{d_Y\})^3] \\ &= \left(\frac{1}{\sigma}\right)^3 \left\{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3\right\} \end{aligned} \tag{7.112}$$

Corollary 7.2.4 (Kurtosis (the fourth standardized moment)).

$$\begin{aligned} Ku[Y|\{d_Y\}] &= E^s[(Y|\{d_Y\})^4] \\ &= \left(\frac{1}{\sigma}\right)^4 \left\{\mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^3\right\} \end{aligned} \tag{7.113}$$

Definition 7.2.17 (Median).

$$Me[Y] = \frac{Q_Y^{(u)}(0.5) + Q_Y^{(\ell)}(0.5)}{2} \tag{7.114}$$

such that

- $Q_Y^{(u)}(0.5)$ *is determined by*

$$\int_0^{\frac{Q_Y^{(u)}(0.5)-a}{b-a}} e^{\beta t+\gamma t^2} dt = \frac{1}{2}\int_0^1 e^{\beta t+\gamma t^2} dt \tag{7.115}$$

- $Q_Y^{(\ell)}(0.5)$ *is determined by*

$$\int_{\frac{Q_Y^{(\ell)}(0.5)-a}{b-a}}^1 e^{\beta t+\gamma t^2} dt = \frac{1}{2}\int_0^1 e^{\beta t+\gamma t^2} dt \tag{7.116}$$

Proposition 7.2.4 (Stochastic entropy).

$$H\left(f_{Y|\{d_Y\}}\right) = \log\left((b-a)\int_0^1 e^{\beta x+\gamma x^2}dx\right) - \beta\mu_1 - \gamma\mu_2 \tag{7.117}$$

Proof of the ***proposition 7.2.4****.* By using (7.91) and the given relations of λ_1 and λ_2, we get the expression of the stochastic entropy as

$$\begin{aligned}
H\left(f_{Y|\{d_Y\}}\right) &= -\left(\lambda_0 + \lambda_1 E[Y|\{d_Y\}] + \lambda_2 E[(Y|\{d_Y\})^2]\right) \\
&= \beta\left(\frac{a}{b-a}\right) - \gamma\left(\frac{a}{b-a}\right)^2 + \log\left((b-a)\int_0^1 e^{\beta x+\gamma x^2}dx\right) \\
&\quad -\left(\frac{\beta}{b-a} - \frac{2a\gamma}{(b-a)^2}\right)\left(a+(b-a)\mu_1\right) \\
&\quad -\frac{\gamma}{(b-a)^2}\left(V[Y|\{d_Y\}] + (E[Y|\{d_Y\}])^2\right) \\
&= \log\left((b-a)\int_0^1 e^{\beta x+\gamma x^2}dx\right) \\
&\quad -\gamma\left(\frac{a}{b-a}\right)^2 - \beta\mu_1 + \frac{2a^2\gamma}{(b-a)^2} + \frac{2a\gamma}{b-a}\mu_1 \\
&\quad -\frac{\gamma}{(b-a)^2}\left[(b-a)^2(\mu_2-\mu_1^2) + \left(a+(b-a)\mu_1\right)^2\right] \\
&= \log\left((b-a)\int_0^1 e^{\beta x+\gamma x^2}dx\right) \\
&\quad -\beta\mu_1 + \frac{a^2\gamma}{(b-a)^2} + \frac{2a\gamma}{b-a}\mu_1 - \gamma(\mu_2-\mu_1^2) \\
&\quad -\gamma\left(\frac{a}{b-a}+\mu_1\right)^2 \\
&= \log\left((b-a)\int_0^1 e^{\beta x+\gamma x^2}dx\right) - \beta\mu_1 - \gamma\mu_2
\end{aligned} \tag{7.118}$$

and this completes the deduction proposed by the **proposition 7.2.4**. □

Chapter 8

Illustrations of standard m. i. probability distributions

In this chapter, we shall **hypothetically** picture a probability distribution in a given situation (though a situation based probability distribution is **generally unknown**) and see how the corresponding minimum information probability distributions **having exactly the same moments** does **fit** into the situation.

This (aforesaid) hypothetical picture is **purely** for the sake of **exemplified illustrations**.

We shall basically confine ourselves to the **continuous** probability distributions for the very simple reason that the graphical representations of the same are clearer.

This is to say, that it is **easier to read the lined density curves** than the **dotted curves** representing probability distributions.

Nevertheless, we shall illustrate both the cases of discrete and continuous probability distributions graphically.

8.1 Examples of minimum information probability distributions

We shall basically analyze the compatibility of given constructed minimum information probability distributions. However, our analysis shall be restricted to the **standard** minimum information probability distributions, as discussed. In other words, as far as our cases are concerned, the availability of the number of moments is at most two. The numerical methods applied for the construction of minimum information probability distributions are already discussed in full details in the second part of this dissertation. So, the **numerical examples**, which are given here, are the results obtained by **running the software programs developed by numerical methods**.

From the **theoretical** point of view, it is absolutely clear, that the betterment of the construction of the probability distributions is directly proportional to the availability of the moments higher moments. But, **in reality**, the exact values of moments are **never known**, but the estimated values of the same. In our illustrated examples, we shall hypothetically assume the theoretical exactness of the moments. Importantly, that the availability of higher moments means higher running times of algorithms as well as higher costs. This is precisely the reason, why we use the minimum information principle. This minimum information principle with regard to uniform, monotonic and uni-extremal is **recapitulated** in the following way:

1. For the construction of a minimum information uniform probability distribution, the minimum information principle says, the range of variability $\mathcal{X}_X$ of the random variable X is necessary to construct the uniform probability distribution of X.

   However, if an additional information with regard to the knowledge of μ_1 is available, it becomes **more than the minimum information** than what was at all necessary. This additional information would be **undesirable**, if, owing to a **possible estimation error**, $\mu_1 \neq \frac{1}{2}$, because the usage of μ_1 in that case **does not enable** the probability distribution of X to be **uniform** anymore.

2. For the construction of a minimum information monotonic probability distribution of X, the minimum information principle says that the following information is necessary:

- $\mathcal{X}_X$
- μ_1

However, if an additional information with regard to the knowledge of μ_2 is available, it becomes again **more than the minimum information** than what was at all necessary. Even in this case, this additional information would be **undesirable**, if, owing to a **possible estimation error**, the condition $\mu_1^2 + \sigma_{X,U}^2 < \mu_2 < \mu_1^2 + \sigma_{X,L}^2$ is not fulfilled, because the usage of μ_2 in that case **does not enable** the probability distribution to be **monotone** anymore.

In this case, if the probability distribution of X has to be specifically **monotone decreasing** or **monotone increasing**, then one must pay a special attention to the different ranges of μ_1.

3. For the construction of a minimum information uni-extremal probability distribution of X, the minimum information principle says that the following information is necessary:

   - $\mathcal{X}_X$
   - μ_1
   - μ_2

   Again, exactly by the same argument, as above, the additional information of μ_3 is **not** utilized.

   In this case, if the probability distribution of X has to be specifically **uni-modal** or **bathtub-shaped**, then one must pay a special attention to the different ranges of μ_2.

Now, we shall illustrate certain minimum information probability distributions in both discrete and continuous cases.

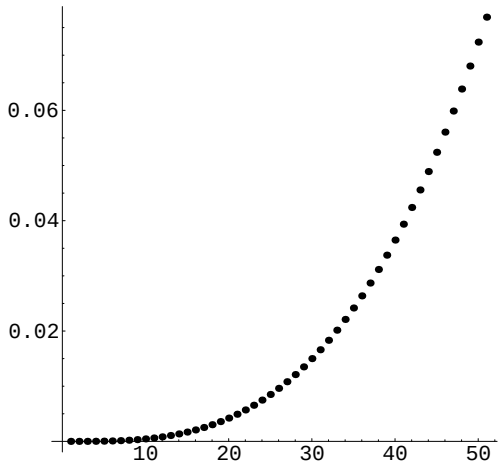
As in discrete cases the overlapping of the two dotted curves cannot be made apparent very easily, the judgement about the goodness of fit of the minimum information probability distributions to the originally existing probability distribution is possible by joining the dots of the probability mass curves to give lined probability mass curves.

8.1.1 Discrete cases

Example 8.1.1 (A monotonic probability distribution). *Let the originally given monotonic probability mass function of the random variable Y be given by*

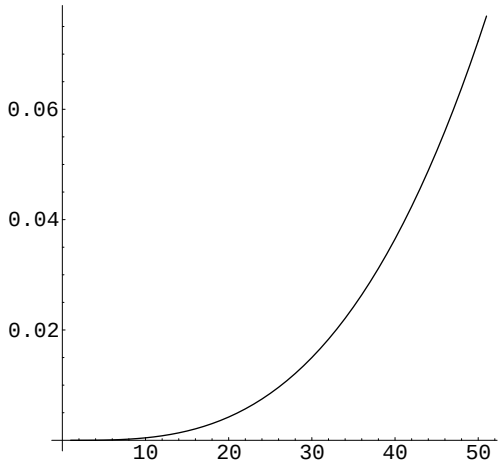
$$f_Y(y) = f_Y(y_j) = \frac{y_j^3}{\sum_{i=0}^{50} i^3}, \quad y_j = j = 0, 1, \ldots, 50 \text{ i.e. } y \in \{y_0, y_1, \ldots, y_{50}\} \tag{8.1}$$

whose dotted probability mass curve is given as



The monotonic dotted probability mass curve $f_Y(y)$

the lined probability mass curve of which is given as



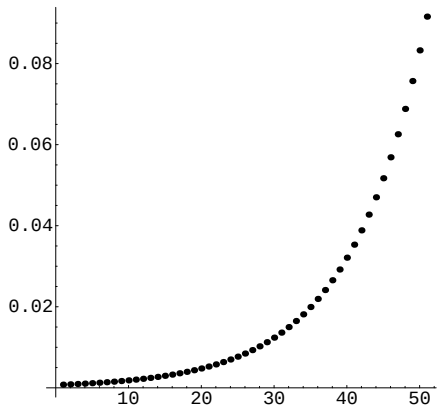
The monotonic lined probability mass curve $f_Y(y)$

and the computed first two moments of the probability distribution are $\mu_Y^{(1)} = 40.3947$ *and* $\mu_Y^{(2)} = 1699.665088$ *(or equivalently* $\sigma_Y^2 = 67.9333$*) respectively.*

Now, the probability mass function of Y*, viz.* $f_{Y|\{(40.3947)\}}(y)$ *giving computed the minimum information probability distribution with subject to* $d_Y = (40.3947)$ *is given as*

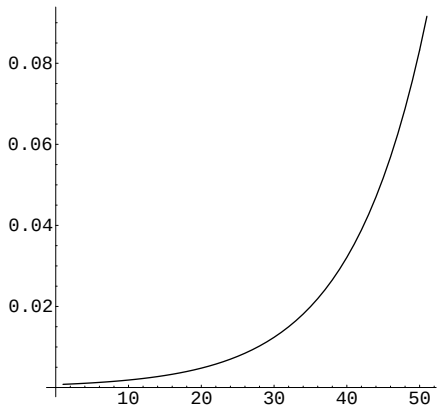
$$f_{Y|\{(40.3947)\}}(y_j) = e^{-7.154126509998741+(0.09527296175581258)*y_j}, \quad y_j = j = 0, 1, \ldots, 50 \tag{8.2}$$

whose dotted probability mass curve is given as



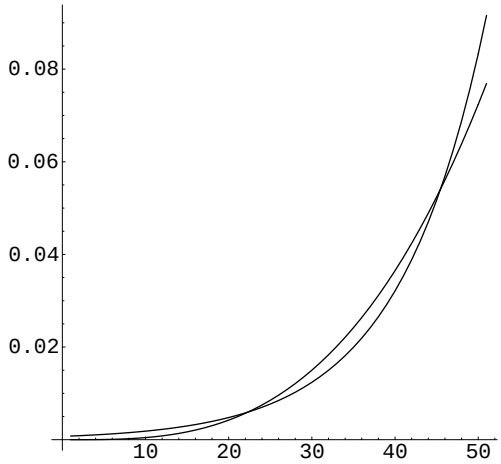
The constructed monotonic dotted probability mass curve $f_{Y|\{(40.3947)\}}(y)$

the lined probability mass curve of which is given as



The constructed monotonic lined probability mass curve $f_{Y|\{(40.3947)\}}(y)$

Clearly, $f_{Y|\{(40.3947)\}}(y)$ *approximates* $f_Y(y)$ *fairly well and this can be easily seen by plotting the corresponding lined mass curves simultaneously as follows*



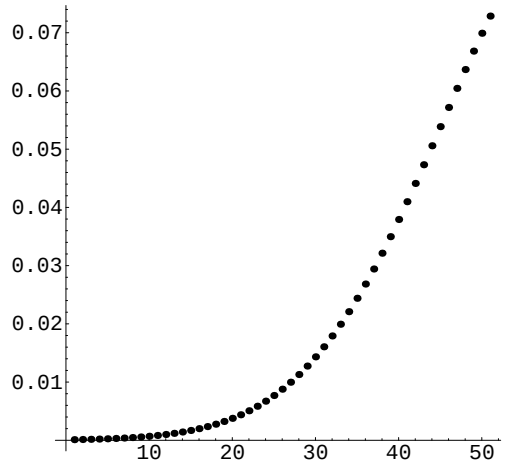
Simultaneous representation of the lined $f_{Y|\{(40.3947)\}}(y)$
and the lined $f_Y(y)$

Moreover, according to Weierstrass, the approximation of $f_Y(y)$ *can still be improved by a further introduction of the knowledge of* $\mu_Y^{(2)}$.

This means, the probability mass function of Y*, viz.* $f_{Y|\{(40.3947,1699.665088)\}}(y)$ *computed with subject to* $d_Y = (40.3947, 1699.665088)$ *is given as*

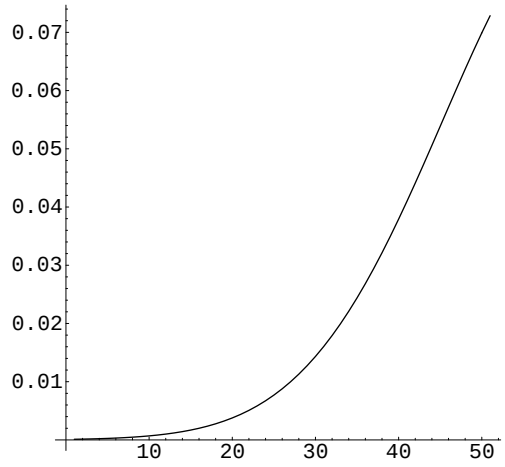
$$\begin{aligned} & f_{Y|\{(40.3947,1699.665088)\}}(y_j) \\ & = e^{-9.110326677351692+(0.2201694347430269)*y_j+(-0.0018069380724544065)*y_j^2}, \\ & \quad y_j = j = 0, 1, \ldots, 50 \end{aligned} \tag{8.3}$$

whose dotted probability mass curve is given as



The constructed monotonic dotted probability mass curve
$f_{Y|\{(40.3947,1699.665088)\}}(y)$

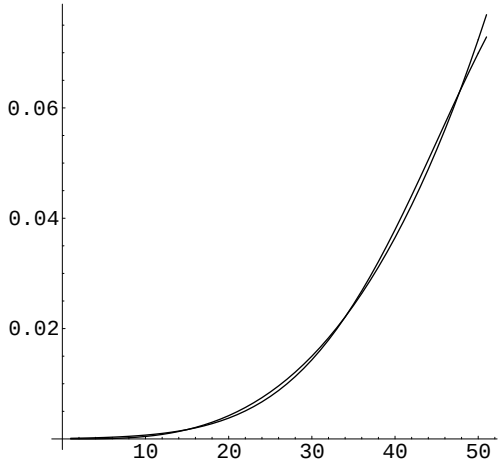
the lined probability mass curve of which is given as



The constructed monotonic lined probability mass curve
$f_{Y|\{(40.3947,1699.665088)\}}(Y)$

Clearly, $f_{Y|\{(40.3947,1699.665088)\}}(y)$ *is a much better approximation of* $f_Y(y)$ *than* $f_{Y|\{(40.3947)\}}(y)$*,*

which can be easily shown by plotting both the lined mass curves of $f_{Y|\{(40.3947,1699.665088)\}}(y)$ *and* $f_Y(y)$ *simultaneously as*



Simultaneous representation of the lined $f_{Y|\{(40.3947,1699.665088)\}}(y)$ and the lined $f_Y(y)$

Thus, from the example 8.1.1, we conclude,

- Every additional knowledge of a higher distributional moment contributes to the improvement of the construction of the probability distribution of Y, which verifies the statement given by Weierstrass.

- Even though the probability mass function $f_{Y|\{(40.3947,1699.665088)\}}(y)$ is a better approximation of the originally given probability mass function $f_Y(y)$ than that of $f_{Y|\{(40.3947)\}}(y)$, according to our definition, $f_{Y|\{(40.3947)\}}(y)$ gives the minimum information probability distribution that does not utilize the knowledge of the second moment $\mu_Y^{(2)} = 1699.665088$.

- Because of the very fact that the equivalence of $\mu_Y^{(2)} = 1699.665088$, namely $\sigma_Y^2 = 67.9333$ ($= \mu_Y^{(2)} - (\mu_Y^{(1)})^2 = 1699.665088 - 40.3947^2$) lies between the limits $\sigma_{Y,U}^2$ and $\sigma_{Y,L}^2$ of monotonicity, i.e.
  $\sigma_{Y,U}^2 = 52.6189 < \sigma_Y^2 = 67.9333 < \sigma_{Y,L}^2 = 108.315$, in this case, $f_{Y|\{(40.3947,1699.665088)\}}(y)$ preserves the monotonic character of the probability distribution.

- Because $N = 51$ is more or less large, the probability distribution of Y may be approximated to a continuous probability distribution and hence the usage of the rules for computing $\sigma^2_{Y,U}$ and $\sigma^2_{Y,L}$ meant for characterizing the probability distribution of Y is more or less appropriate.

- The entropy of the uniform probability distribution with the same support is $\log 51 = 3.93183$

- The entropy of the probability distribution given by the density $f_{Y|\{(40.3947)\}}$ is given as

$$\begin{aligned} &- (-7.154126509998741 + 0.09527296175581258\mu_Y^{(1)}) \\ &= 3.3056 \end{aligned}$$

  which is clearly less than $\log 51$.

- Lastly, the entropy of the probability distribution given by the density $f_{Y|\{(40.3947,1699.665088)\}}$ is given as

$$\begin{aligned} &- (-9.110326677351692 + 0.2201694347430269\mu_Y^{(1)} \\ &- 0.0018069380724544065\mu_Y^{(2)}) = 3.28784 \end{aligned}$$

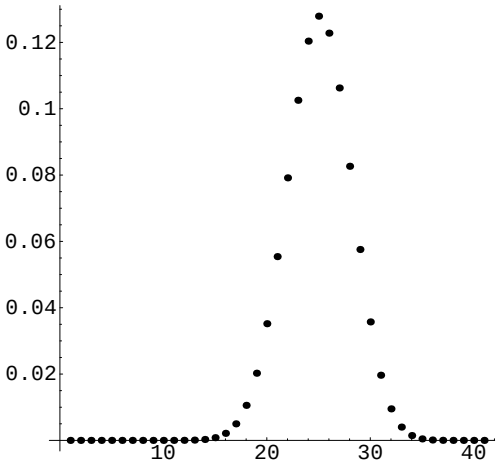
  which is clearly less than $\log 51$.

Example 8.1.2 (A binomial probability distribution). *The graphically represented binomial distribution is a very commonly known bell- shaped dotted figure and due to its bell- shapeliness, it is an uni- extremal and in fact an uni- modal probability distribution. Thus, the construction of the fitting appropriate minimum information probability distribution understandably necessitates the availability of at least two moments.*

Let the binomial probability mass function of the random variable Y be given by

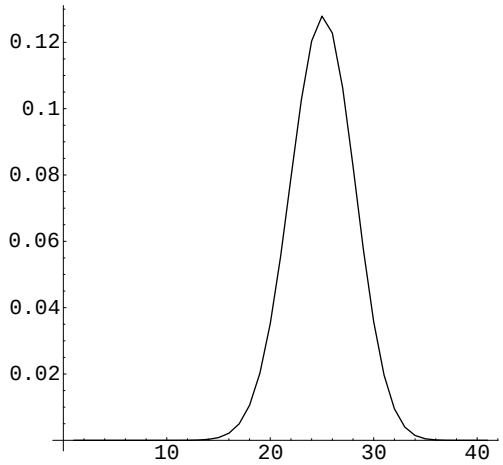
$$\begin{aligned} f_Y(y) = f_Y(y_j) = \tbinom{40}{j}(0.6)^j\,(0.4)^{40-j}, \quad y_j = j = 0, 1, \ldots, 40 \\ i.e.\ y \in \{y_0, y_1, \ldots, y_{40}\} \end{aligned} \tag{8.4}$$

whose dotted probability mass curve is given as



The binomial dotted probability mass curve $f_Y(y)$

the lined probability mass curve of which is given as



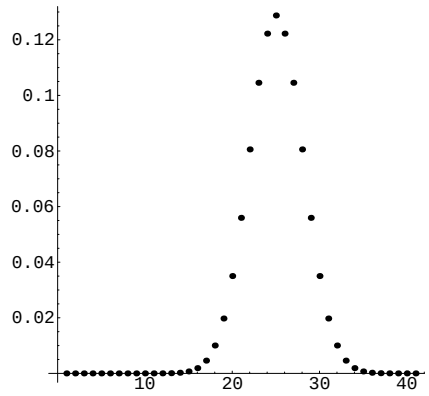
The binomial lined probability mass curve $f_Y(y)$

and the computed moments of the probability distribution are $\mu_Y^{(1)} = 24$ *and* $\mu_Y^{(2)} = 585.6$ *(or equivalently* $\sigma_Y^2 = 9.6$*) respectively.*

Now, the probability mass function of Y*, viz.* $f_{Y|\{(24,585.6)\}}(y)$ *giving the computed minimum information probability distribution with subject to* $d_Y = (24, 585.6)$ *is given as*

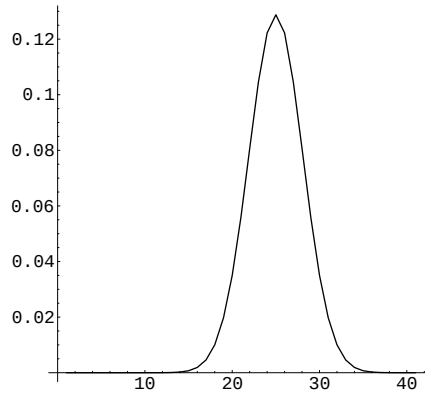
$$\begin{aligned} & f_{Y|\{(24,585.6)\}}(y_j) \\ & = e^{-32.04982003800298+(2.5)*y_j+(-0.05208333333333336)*y_j^2}, \\ & \quad y_j = j = 0, 1, \ldots, 40 \end{aligned} \tag{8.5}$$

whose dotted probability mass curve is given as



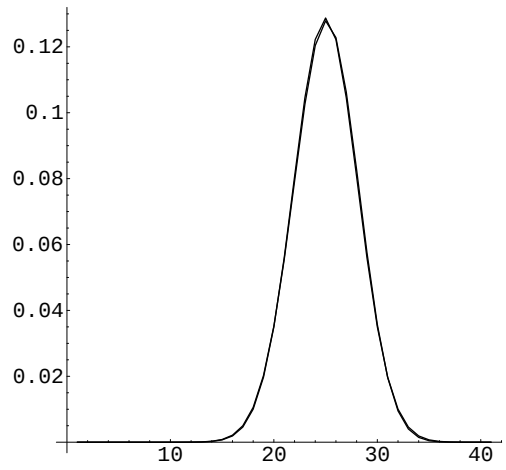
The constructed binomial dotted probability mass curve $f_{Y|\{(24,585.6)\}}(y)$

the lined probability mass curve of which is given as



The constructed binomial lined probability mass curve
$f_{Y|\{(24,585.6)\}}(y)$

Clearly, $f_{Y|\{(24,585.6)\}}(y)$ approximates $f_Y(y)$ excellently, which can be easily shown by plotting both the lined mass curves of $f_{Y|\{(24,585.6)\}}(y)$ and $f_Y(y)$ simultaneously as



Simultaneous representation of the lined $f_{Y|\{(24,585.6)\}}(y)$
and the lined $f_Y(y)$

Thus, from the example 8.1.2, we conclude,

- Because of the very fact that the equivalence of $\mu_Y^{(2)} = 585.6$, namely $\sigma_Y^2 = 9.6$ ($= \mu_Y^{(2)} - (\mu_Y^{(1)})^2 = 585.5 - 24^2$) lies well below $\sigma_{Y,U}^2 = 113.05$ i.e. $\sigma_Y^2 = 9.6 < \sigma_{Y,U}^2 = 113.05$, in this case, $f_{Y|\{(24,585.6)\}}(y)$ preserves the uni- modal character of the probability distribution.

- Because $N = 41$ is more or less large, the probability distribution of Y may be approximated to a continuous probability distribution and hence the usage of the rules for computing $\sigma_{Y,U}^2$ and $\sigma_{Y,L}^2$ meant for characterizing the probability distribution of Y is more or less appropriate.

- The entropy of the uniform probability distribution with the same support is $\log 41 = 3.71357$

- Lastly, the entropy of the probability distribution given by the density

$f_{Y|\{(24,585.6)\}}$ is given as

$$\begin{aligned}&-(-32.04982003800298+2.5\mu_Y^{(1)}\\&-0.052083333333333336\mu_Y^{(2)})=2.54982\end{aligned}$$

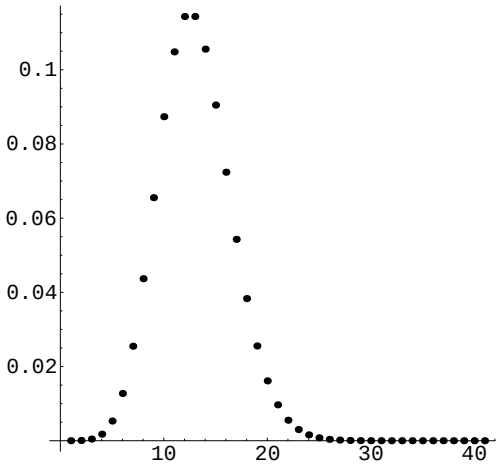
which is clearly less than $\log 41$.

Example 8.1.3 (A truncated poisson distribution). *Even graphically represented truncated poisson distribution is a commonly known bell- shaped dotted figure and due to its bell- shapeliness, the fitting of an appropriate minimum information probability distribution understandably necessitates the availability of at least two moments as well.*

Let the truncated poisson probability mass function of the random variable Y be given by

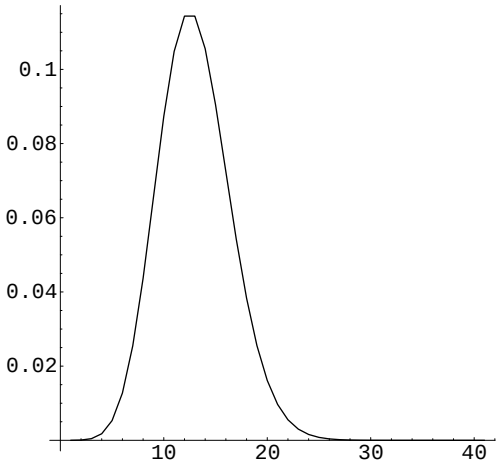
$$f_Y(y) = f_Y(y_j) = \frac{\frac{12^j}{j!}}{\sum_{i=0}^{40} \frac{12^i}{i!}}, \quad y_j = j = 0, 1, \ldots, 40$$
$$i.e.\ y \in \{y_0, y_1, \ldots, y_{40}\} \tag{8.6}$$

whose dotted probability mass curve is given as



The truncated poisson dotted probability mass curve $f_Y(y)$

the lined probability mass curve of which is given as



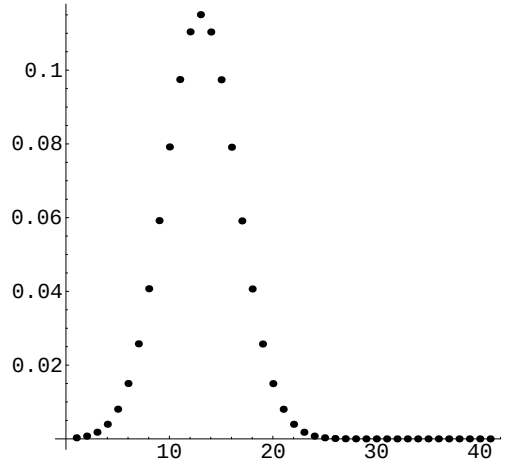
The truncated poisson lined probability mass curve $f_Y(y)$

and the computed moments of the probability distribution are $\mu_Y^{(1)} = 12$ *and* $\mu_Y^{(2)} = 156$ *(or equivalently* $\sigma_Y^2 = 12$*) respectively.*

Now, the probability density function of Y*, viz.* $f_{Y|\{(12,156)\}}(y)$ *giving the computed minimum information probability distribution with subject to* $d_Y = (12, 156)$ *is given as*

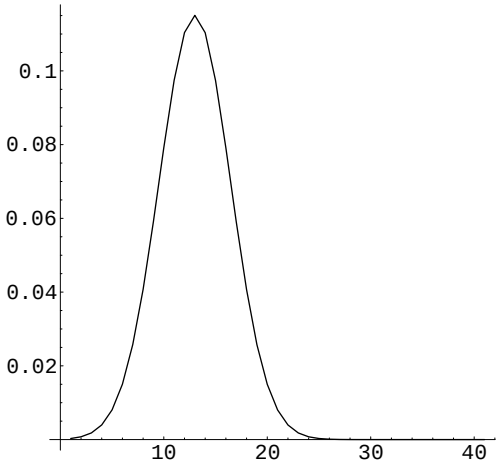
$$\begin{aligned} & f_{Y|\{(12,156)\}}(y_j) \\ & = e^{-8.147693845476917+(0.9977335224603812)*y_j+(-0.04157918112705276)*y_j^2}, \\ & \quad y_j = j = 0, 1, \ldots, 40 \end{aligned} \tag{8.7}$$

whose dotted probability mass curve is given as



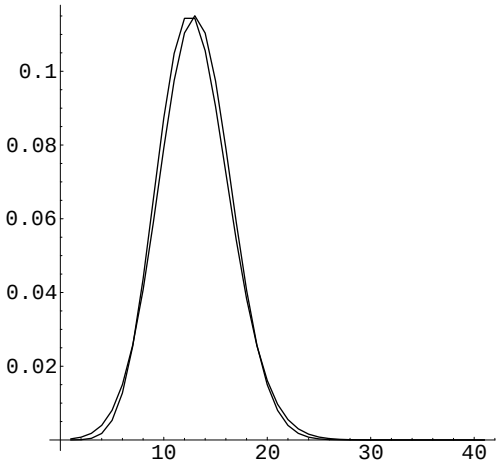
The constructed truncated poisson dotted probability mass curve $f_{Y|\{(12,156)\}}(y)$

the lined probability mass curve of which is given as



The constructed truncated poisson lined probability mass curve
$f_{Y|\{(12,156)\}}(y)$

Clearly, $f_{Y|\{(12,156)\}}(y)$ *approximates* $f_Y(y)$ *wonderfully, which can be easily shown by plotting both the lined mass curves of* $f_{Y|\{(12,156)\}}(y)$ *and* $f_Y(y)$ *simultaneously as*



Simultaneous representation of the lined $f_{Y|\{(12,156)\}}(y)$
and the lined $f_Y(y)$

Thus, from the example 8.1.3, we conclude,

- Because of the very fact that the equivalence of $\mu_Y^{(2)} = 156$, namely $\sigma_Y^2 = 12$ ($= \mu_Y^{(2)} - (\mu_Y^{(1)})^2 = 156 - 12^2$) lies well below $\sigma_{Y,U}^2 = 76.8806$ i.e. $\sigma_Y^2 = 12 < \sigma_{Y,U}^2 = 76.8806$, in this case, $f_{Y|\{(12,156)\}}(y)$ preserves the uni- modal character of the probability distribution.

- Because $N = 41$ is more or less large, the probability distribution of Y may be approximated to a continuous probability distribution and hence the usage of the rules for computing $\sigma_{Y,U}^2$ and $\sigma_{Y,L}^2$ meant for characterizing the probability distribution of Y is more or less appropriate.

- The entropy of the uniform probability distribution with the same support is $\log 41 = 3.71357$

- Lastly, the entropy of the probability distribution given by the density $f_{Y|\{(12,156)\}}$ is given as

$$\begin{aligned} &- (-8.147693845476917 + 0.9977335224603812\mu_Y^{(1)} \\ &- 0.04157918112705276\mu_Y^{(2)}) = 2.66124 \end{aligned}$$

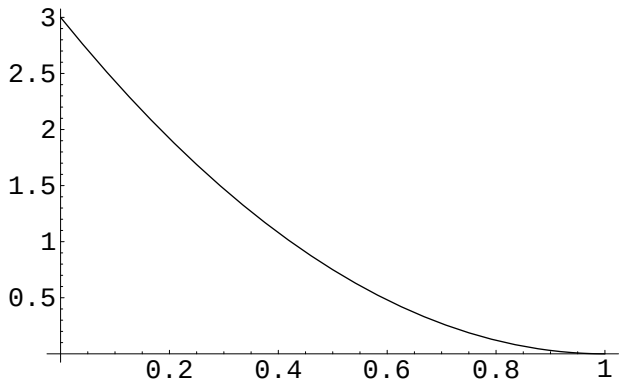
  which is clearly less than $\log 41$.

8.1.2 Continuous cases

Example 8.1.4 (A monotonic probability distribution). *Let the monotonic probability density function of the random variable X be given by*

$$f_X(x) = 3(1-x)^2, \quad 0 \leq x \leq 1 \tag{8.8}$$

whose the probability density curve, which is a monotonic curve, is given as



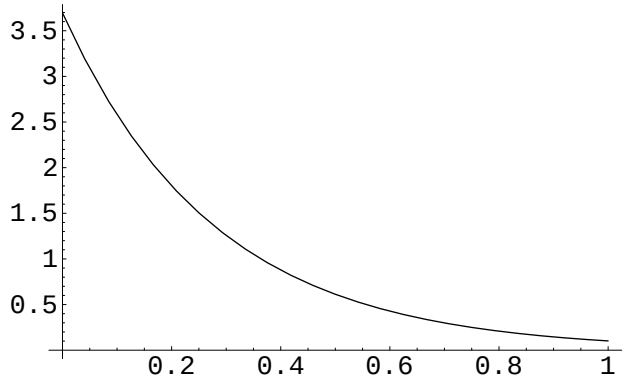
The monotonic probability density curve $f_X(x)$

and the computed first two moments of the probability distribution are $\mu_1 = 0.25$ and $\mu_2 = 0.1$ respectively.

Now, the probability density function of X, viz. $f_{X|\{(0.25)\}}(x)$ giving computed the minimum information probability distribution with subject to $d = (0.25)$ is given as

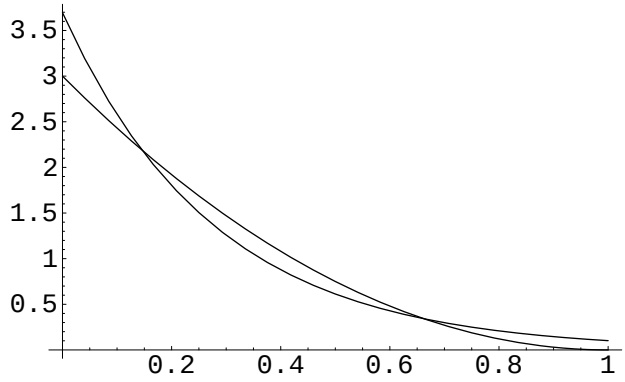
$$f_{X|\{(0.25)\}}(x) = 3.6951339770855713e^{-3.593511969447428x}, 0 \leq x \leq 1 \tag{8.9}$$

whose density curve is given as



The constructed monotonic probability density curve $f_{X|\{(0.25)\}}(x)$

Clearly, $f_{X|\{(0.25)\}}(x)$ is a fairly good approximation of $f_X(x)$, which can be shown by plotting both of them simultaneously as



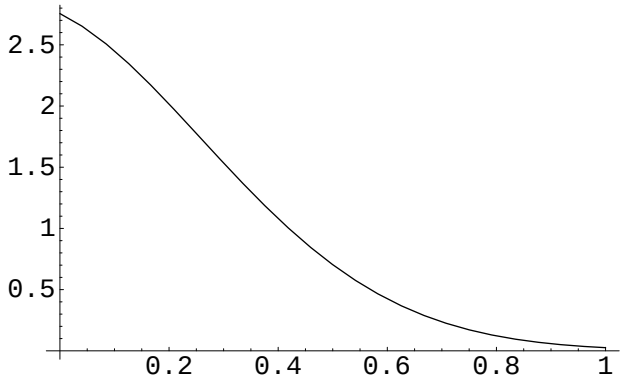
Simultaneous representation of $f_{X|\{(0.25)\}}(x)$ and $f_X(x)$

Moreover, the approximation of $f_X(x)$ can still be improved by a further introduction of the knowledge of $E[X^2]$, in case the cost of the acquirement of this knowledge is affordable.

This means, the probability density function of X, viz. $f_{X|\{(0.25,0.1)\}}(x)$ computed by means of the maximum entropy principle with subject to the known $d = (\mu_1, \mu_2) = (0.25, 0.1)$ is given as

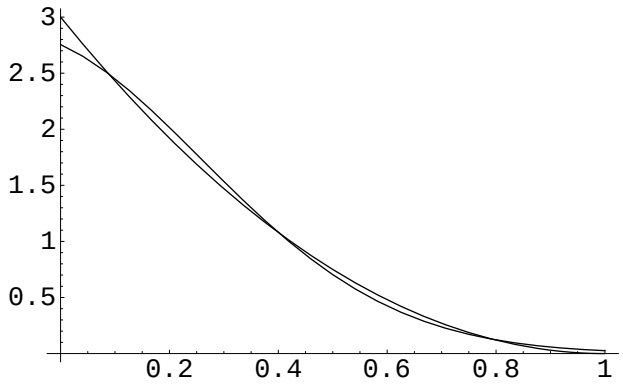
$$f_{X|\{(0.25,0.1)\}}(x) = \quad 2.7553146976923863e^{-0.783542712820377x-3.892192210257011x^2}, \quad 0 \leq x \leq 1 \tag{8.10}$$

whose density curve is given as



The constructed monotonic probability density curve $f_{X|\{(0.25,0.1)\}}(x)$

Clearly, $f_{X|\{(0.25,0.1)\}}(x)$ is a better approximation of $f_X(x)$ than $f_{X|\{(0.25)\}}(x)$, which can be easily shown by plotting both of them simultaneously as



Simultaneous representation of $f_{X|\{(0.25,0.1)\}}(x)$ and $f_X(x)$

Thus, from the example 8.1.4, we conclude,

- Every additional knowledge of a higher distributional moment contributes to the improvement of the construction of the probability distribution of X, which verifies the statement given by Weierstrass.
- Even though the probability mass function $f_{X|\{(0.25,0.1)\}}(x)$ is a better approximation of the originally given probability density function $f_X(x)$ than that of $f_{X|\{(0.25)\}}(x)$, according to our definition, $f_{X|\{(0.25)\}}(x)$ gives the minimum information probability distribution that does not utilize the knowledge of the second moment $\mu_2 = 0.1$.

- Because of the very fact that the equivalence of $\mu_2 = 0.1$, namely $\sigma^2 = 0.0375$ ($= \mu_2 - \mu_1^2 = 0.1 - 0.25^2$) lies between the limits $\sigma^2_{X,U}$ and $\sigma^2_{X,L}$ of monotonicity, i.e.
  $\sigma^2_{X,U} = 0.0350331 < \sigma^2 = 0.0375 < \sigma^2_{X,L} = 0.0594589$, in this case, $f_{X|\{(0.25,0.1)\}}(x)$ preserves the monotonic character of the probability distribution.

- The entropy of the probability distribution given by the density $f_{X|\{(0.25)\}}$ is given as

$$\begin{aligned} &- (1.3070168127646282 - 3.593511969447428\mu_1) \\ &= -0.408639 \end{aligned}$$

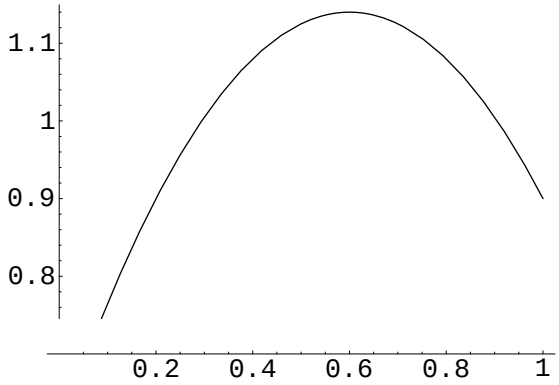
- Lastly, the entropy of the probability distribution given by the density $f_{X|\{(0.25,0.1)\}}$ is given as

$$\begin{aligned} &- (1.013531663918902 - 0.783542712820377\mu_1 \\ &- 3.892192210257011\mu_2) = -0.428427 \end{aligned}$$

Example 8.1.5 (A parabolic uni-extremal probability distribution). *Let the uni- extremal probability density function of the random variable X be given by*

$$f_X(x) = 0.6(1 + 3x - 2.5x^2), \quad 0 \leq x \leq 1 \tag{8.11}$$

whose probability density curve, which is a parabolic bell- shaped curve, is given as



The parabolic probability density curve $f_X(x)$

and the computed first two moments of the probability distribution are $\mu_1 = 0.525$ *and* $\mu_2 = 0.35$ *respectively.*

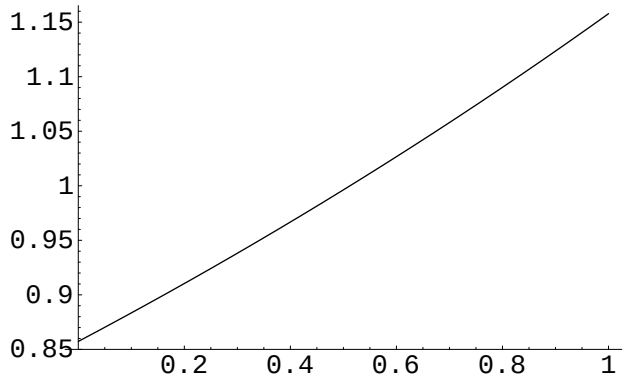
As a matter of fact, due to the bell- shapeliness of the density curve, it is an uni- extremal and in fact an uni- modal probability distribution. Thus, the construction of the fitting appropriate minimum information probability distribution understandably necessitates the availability of at least two moments.

Now, even if the knowledge of μ_1 only (i.e. without the knowledge of μ_2) is not enough to construct the probability density function of X, we shall see here, how the constructed probability of X with subject to $d = (\mu_1) = (0.525)$ behaves:

The probability density function of X viz. $f_{X|\{(0.525)\}}(x)$ computed with subject to $d = (0.525)$ is given as

$$f_{X|\{(0.525)\}}(x) = 0.8572857448518301e^{0.30045106346983064x}, 0 \leq x \leq 1 \tag{8.12}$$

whose density curve is given as



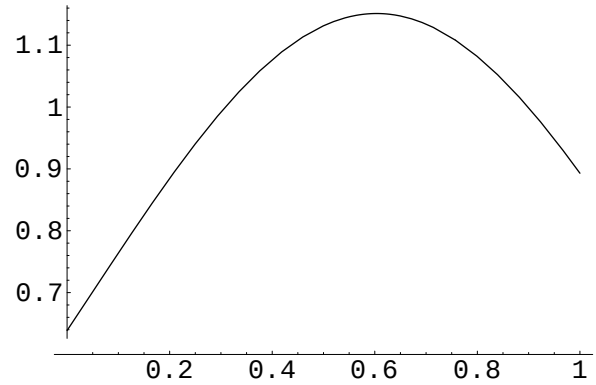
The constructed probability density curve $f_{X|\{(0.525)\}}(x)$

Clearly, $f_{X|\{(0.525)\}}(x)$ is undoubtedly a bad approximation of $f_X(x)$ and therefore an introduction of the further knowledge of μ_2 is imperatively necessary.

Therefore, the probability density function of X, viz. $f_{X|\{(0.525,0.35)\}}(x)$ giving the computed minimum information probability distribution with subject to $d = (\mu_1, \mu_2) = (0.525, 0.35)$ is given as

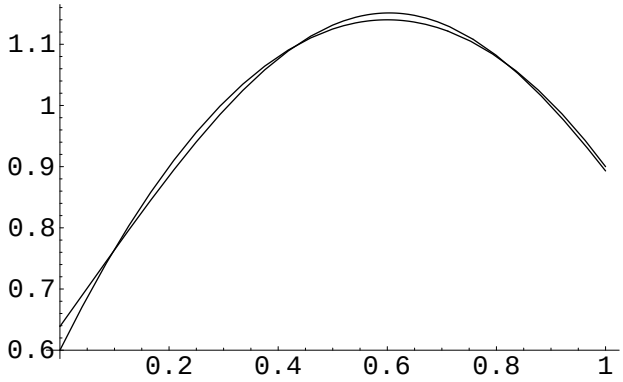
$$\begin{aligned} &f_{X|\{(0.525,0.35)\}}(x) \\ &= 0.6387466516222691e^{1.951702233911951x - 1.6165863220869872x^2}, 0 \leq x \leq 1 \end{aligned} \tag{8.13}$$

whose density curve is given as



The constructed parabolic probability density curve $f_{X|\{(0.525,0.35)\}}(x)$

Clearly, $f_{X|\{(0.525,0.35)\}}(x)$ is undoubtedly a vastly better approximation of $f_X(x)$ than $f_{X|\{(0.525)\}}(x)$, which can be easily shown by plotting both of them simultaneously as



Simultaneous representation of $f_{X|\{(0.525,0.35)\}}(x)$ and $f_X(x)$

Thus, from the example 8.1.5, we conclude,

- A minimum number of distributional moments of X is direly necessary for the purpose of an acceptable construction of it's probability distribution. In this example, we have demonstrated the very fact that the minimum information probability distribution demands minimum number of distributional moments.

- Because of the very fact that the equivalence of $\mu_2 = 0.35$, namely $\sigma^2 = 0.074375$ ($= \mu_2 - \mu_1^2 = 0.35 - 0.525^2$) lies well below $\sigma^2_{X,U} = 0.081274$ i.e. $\sigma^2 = 0.074375 < \sigma^2_{X,U} = 0.081274$, in this case, $f_{X|\{(0.525,0.35)\}}(x)$ preserves the uni-modal character of the probability distribution.

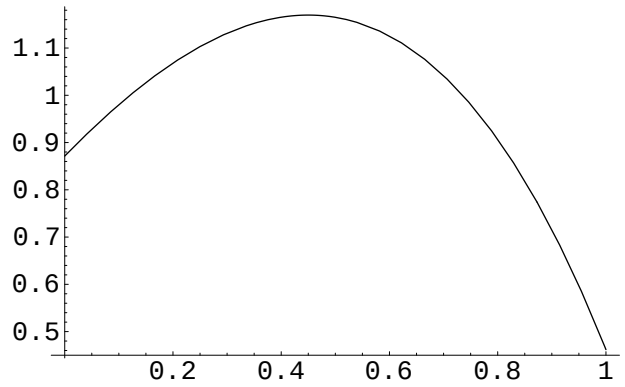
- Lastly, the entropy of the probability distribution given by the density $f_{X|\{(0.525,0.35)\}}$ is given as

$$-(-0.44824737955660204 + 1.951702233911951\mu_1 - 1.6165863220869872\mu_2) = -0.0105911$$

Example 8.1.6 (A non-parabolic uni-extremal probability distribution). *Let the uni-extremal probability density function of the random variable X be given by*

$$f_X(x) = \frac{4}{459}(100 + 135x - 96x^2 - 71x^3 - 15x^4), \quad 0 \leq x \leq 1 \qquad (8.14)$$

whose probability density curve, which is a non-parabolic bell-shaped curve, is given as



The non-parabolic probability density curve $f_X(x)$

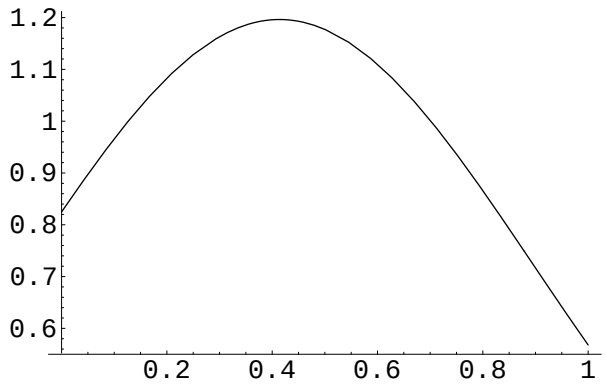
and the computed first two moments of the probability distribution are $\mu_1 = 0.473203$ *and* $\mu_2 = 0.295487$ *respectively.*

As a matter of fact, due to the bell- shapeliness of the density curve, it is an uni- extremal and in fact an uni- modal probability distribution. Thus, the construction of the fitting appropriate minimum information probability distribution understandably necessitates the availability of at least two moments.

In our immediately preceding example we have illustratively demonstrated that at least two moments of X are necessary for the construction of the probability distribution of X. Exactly the same is the case, in this very illustrated example as well. Keeping this in mind, the probability density function of X, viz. $f_{X|\{(0.473203,0.295487)\}}(x)$ giving the computed minimum information probability distribution with subject to $d = (0.473203, 0.295487)$ is given as

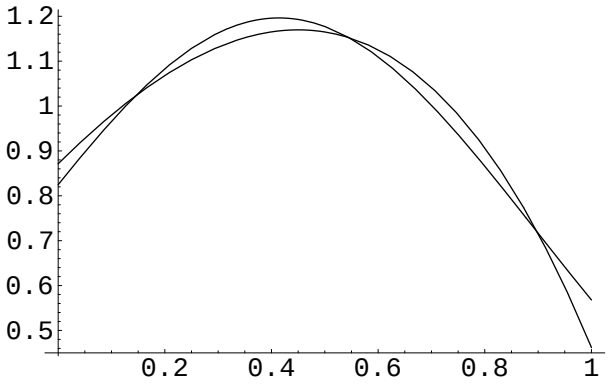
$$f_{X|\{(0.473203,0.295487)\}}(x)$$
$$= 0.8248284075199338e^{1.7961285412633676x - 2.1692651859053607x^2}, 0 \leq x \leq 1 \qquad (8.15)$$

whose density curve is given as



The constructed non-parabolic probability density curve
$f_{X|\{(0.473203,0.295487)\}}(x)$

Clearly, $f_{X|\{(0.473203,0.295487)\}}(x)$ *is undoubtedly a very good approximation of* $f_X(x)$*, which can be easily shown by plotting both of them simultaneously as*



Simultaneous representation of $f_{X|\{(0.473203,0.295487)\}}(x)$ and $f_X(x)$

Thus, from the example 8.1.6, we conclude,

- Because of the very fact that the equivalence of $\mu_2 = 0.295487$, namely $\sigma^2 = 0.0715659$ ($= \mu_2 - \mu_1^2 = 0.295487 - 0.473203^2$) lies well below $\sigma^2_{X,U} = 0.081096$ i.e.
  $\sigma^2 = 0.0715659 < \sigma^2_{X,U} = 0.081096$, in this case, $f_{X|\{(0.473203,0.295487)\}}(x)$ preserves the uni- modal character of the probability distribution.

- Lastly, the entropy of the probability distribution given by the density $f_{X|\{(0.473203,0.295487)\}}$ is given as

$$\begin{aligned}&-(-0.19257990516548862+1.7961285412633676\mu_1\\&-2.1692651859053607\mu_2)=-0.016363\end{aligned}$$

Note: In the next few examples we shall handle the well known beta distribution for different parametric values. We shall denote the parameters of the beta distribution by k_1 and k_2.

Example 8.1.7 (A beta distribution with $k_1 < 1$ and $k_2 < 1$). *With subject to $k_1 = 0.49$ and $k_2 = 0.66$, the beta distribution as a uni-extremal bathtub-shaped probability density function of the random variable X is given by*

$$f_X(x) = \frac{0.377735}{(1-x)^{0.34}x^{0.51}}, \quad 0 \leq x \leq 1 \tag{8.16}$$

and the computed first two moments of the probability distribution are $\mu_1 = 0.426087$ and $\mu_2 = 0.295288$ respectively.

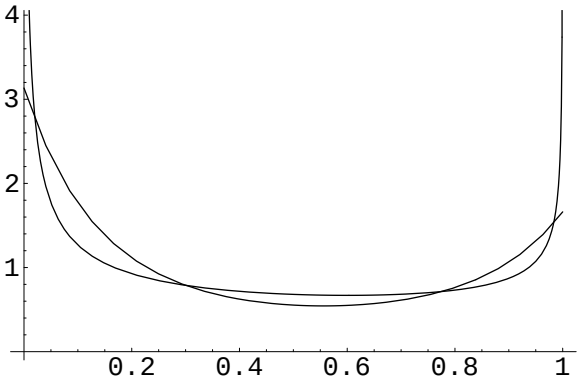
As a matter of fact, due to the bathtub- shapeliness of the density curve, it is an uni- extremal probability distribution.

Thus, the construction of the fitting appropriate minimum information probability distribution understandably necessitates the availability of at least two moments.

Keeping this in mind, the probability density function of X, viz. $f_{X|\{(0.426087,0.295288)\}}(x)$ giving the computed minimum information probability distribution with subject to $d = (0.426087, 0.295288)$ is given as

$$\begin{aligned} &f_{X|\{(0.426087,0.295288)\}}(x) \\ &= e^{1.1420835823040651-6.299235717069621x+5.664887467832139x^2}, 0 \leq x \leq 1 \end{aligned} \tag{8.17}$$

Clearly, $f_{X|\{(0.426087,0.295288)\}}(x)$ is undoubtedly a very good approximation of $f_X(x)$, which can be easily shown by plotting both of them simultaneously as



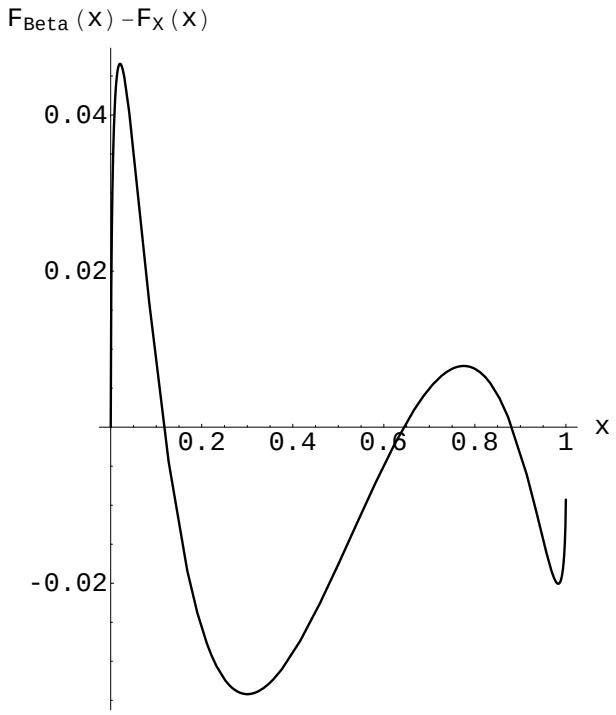
Simultaneous representation of $f_{X|\{(0.426087,0.295288)\}}(x)$ and $f_X(x)$

The very fact that, unlike the curve $f_{X|\{(0.426087,0.295288)\}}(x)$*, both the left and the right branches of the curve* $f_X(x)$ *shoot infinitely upwards, is purely due to the singularities of the curve* $f_X(x)$ *at both the extreme points* $x = 0$ *and* $x = 1$*.*

Thus, we conclude,

- *Because of the very fact that the equivalence of* $\mu_2 = 0.295288$*, namely* $\sigma^2 = 0.0113738$ *(* $= \mu_2 - \mu_1^2 = 0.295288 - 0.426087^2$*) lies well above* $\sigma^2_{X,L} = 0.0847275$ *i.e.*
  $\sigma^2 = 0.113738 > \sigma^2_{X,L} = 0.0847275$*,* $f_{X|\{(0.426087,0.295288)\}}(x)$ *preserves the bathtub shapeliness character of the probability distribution in this case.*

- *Moreover, the differences in the values of the distribution functions with regard to the two above probability distributions, namely the differences* $F_{Beta}(x) - F_X(x)$ *for different values of* $x \in (0, 1)$*, such that* $F_{Beta}(x) = \int_0^x f_X(t)dt$ *and* $F_X(x) = \int_0^x f_{X|\{(0.426087,0.295288)\}}(t)dt$*, are plotted simultaneously as follows:*



Representation of the difference of $F_{X|\{(0.426087,0.295288)\}}(x)$ and $F_X(x)$

which evidently shows that the aforesaid maximum difference is roughly 4% and therefore $f_{X|\{(0.426087,0.295288)\}}(x)$ *approximates* $f_X(x)$ *well enough.*

- *Lastly, the entropy of the probability distribution given by the density* $f_{X|\{(0.426087,0.295288)\}}$ *is given as*

$$-(1.1420835823040651 - 6.299235717069621\mu_1 + 5.664887467832139\mu_2) = -0.130834$$

Example 8.1.8 (A beta distribution with $k_1 < 1$ and $k_2 > 1$). *With subject to $k_1 = 0.83$ and $k_2 = 3.29$, the beta distribution as a strictly monotone decreasing probability density function of the random variable X is given by*

$$f_X(x) = \frac{2.32477(1-x)^{2.29}}{x^{0.17}}, \quad 0 \leq x \leq 1 \tag{8.18}$$

and the computed first two moments of the probability distribution are $\mu_1 = 0.201456$ and $\mu_2 = 0.0720049$ respectively.

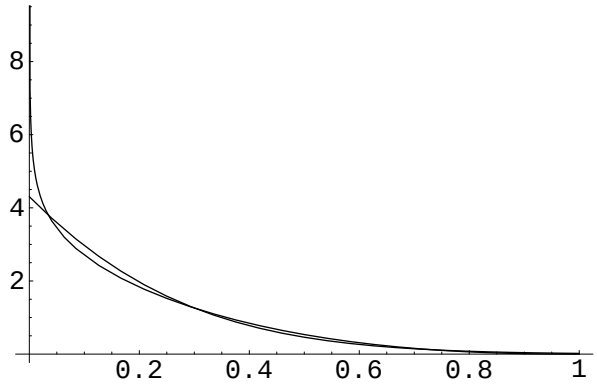
As a matter of fact, due to the monotonic decreasing character of the density curve, it is a monotonic probability distribution.

Thus, the construction of the fitting appropriate minimum information probability distribution understandably necessitates the availability of at least one moment. But for the sake of a better accuracy, we shall make use of two moments.

Keeping this in mind, the probability density function of X, viz. $f_{X|\{(0.201456,0.0720049)\}}(x)$ giving the computed minimum information probability distribution with subject to $d = (0.201456, 0.0720049)$ is given as

$$\begin{aligned} &f_{X|\{(0.201456,0.0720049)\}}(x) \\ &= e^{1.4597752077817678 - 3.5341373793580133x - 1.864903007677731x^2}, 0 \leq x \leq 1 \end{aligned} \tag{8.19}$$

Clearly, $f_{X|\{(0.201456,0.0720049)\}}(x)$ is undoubtedly a very good approximation of $f_X(x)$, which can be easily shown by plotting both of them simultaneously as



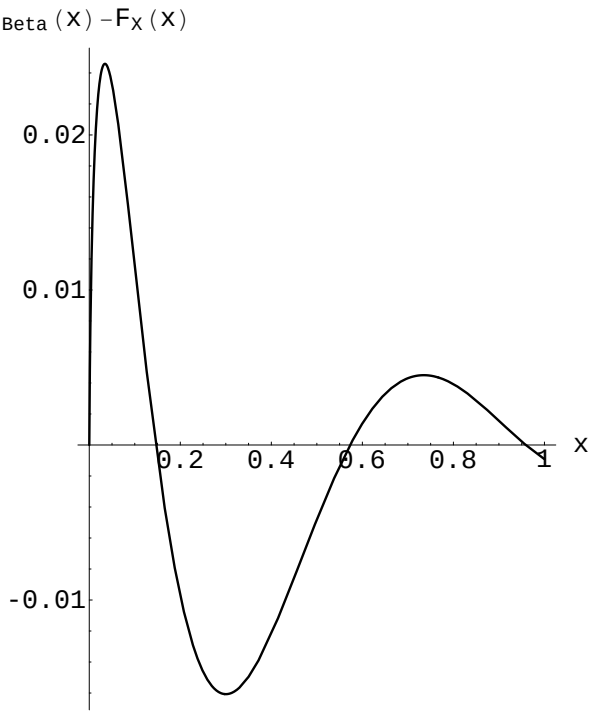
Simultaneous representation of $f_{X|\{(0.201456,0.0720049)\}}(x)$ and $f_X(x)$

The very fact that, unlike the curve $f_{X|\{(0.201456,0.0720049)\}}(x)$*, the left branch of the curve* $f_X(x)$ *shoots infinitely upwards, is purely due to the singularity of the curve* $f_X(x)$ *at the extreme point* $x = 0$*.*

Thus, we conclude,

- *Because of the very fact that the equivalence of* $\mu_2 = 0.0720049$*, namely* $\sigma^2 = 0.0314204$ *(* $= \mu_2 - \mu_1^2 = 0.0720049 - 0.201456^2$*) lies in between* $\sigma^2_{X,U} = 0.0231267$ *and* $\sigma^2_{X,L} = 0.0461303$ *i.e.* $\sigma^2_{X,U} = 0.0231267 < \sigma^2 = 0.0314204 < \sigma^2_{X,L} = 0.0461303$*,* $f_{X|\{(0.201456,0.0720049)\}}(x)$ *preserves the monotonic character of the probability distribution in this case.*

- *Moreover, the differences in the values of the distribution functions with regard to the two above probability distributions, namely the differences* $F_{Beta}(x) - F_X(x)$ *for different values of* $x \in (0, 1)$*, such that* $F_{Beta}(x) = \int_0^x f_X(t)dt$ *and* $F_X(x) = \int_0^x f_{X|\{(0.201456,0.0720049)\}}(t)dt$*, are plotted simultaneously as follows:*



Representation of the difference of $F_{X|\{(0.201456,0.0720049)\}}(x)$ and $F_X(x)$

which evidently shows that the aforesaid maximum difference is roughly 2% and therefore $f_{X|\{(0.201456,0.0720049)\}}(x)$ *approximates* $f_X(x)$ *well enough.*

- *Lastly, the entropy of the probability distribution given by the density* $f_{X|\{(0.201456,0.0720049)\}}$ *is given as*

$$-(1.4597752077817678 - 3.5341373793580133\mu_1 - 1.864903007677731\mu_2) = -0.61352$$

Example 8.1.9 (A beta distribution with $k_1 > 1$ and $k_2 < 1$). *With subject to $k_1 = 2.15$ and $k_2 = 0.72$, the beta distribution as a strictly monotone increasing probability density function of the random variable X is given by*

$$f_X(x) = \frac{1.30879x^{1.15}}{(1-x)^{0.28}}, \quad 0 \leq x \leq 1 \tag{8.20}$$

and the computed first two moments of the probability distribution are $\mu_1 = 0.749129$ and $\mu_2 = 0.609756$ respectively.

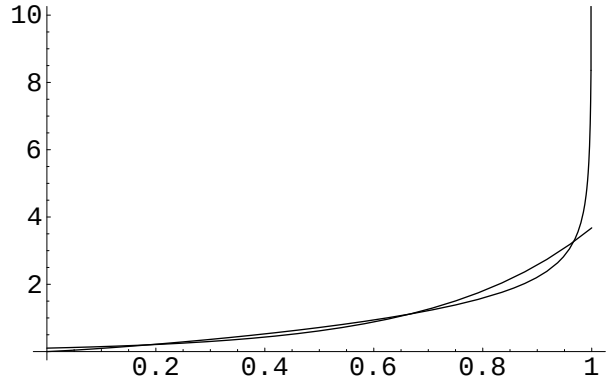
As a matter of fact, due to the monotonic increasing character of the density curve, it is a monotonic probability distribution.

Thus, the construction of the fitting appropriate minimum information probability distribution understandably necessitates the availability of at least one moment. But for the sake of a better accuracy, we shall make use of two moments.

Keeping this in mind, the probability density function of X, viz. $f_{X|\{(0.749129,0.609756)\}}(x)$ giving the computed minimum information probability distribution with subject to $d = (0.749129, 0.609756)$ is given as

$$\begin{aligned} &f_{X|\{(0.749129,0.609756)\}}(x) \\ &= e^{-2.2750656018479605+3.5834317748920568x-0.006347275972631056x^2}, 0 \leq x \leq 1 \end{aligned} \tag{8.21}$$

Clearly, $f_{X|\{(0.749129,0.609756)\}}(x)$ is undoubtedly a very good approximation of $f_X(x)$, which can be easily shown by plotting both of them simultaneously as



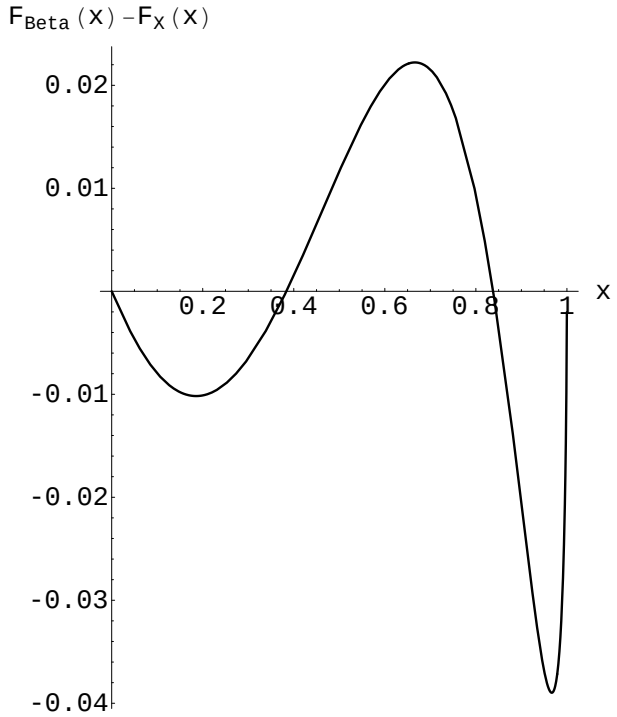
Simultaneous representation of $f_{X|\{(0.749129,0.609756)\}}(x)$ and $f_X(x)$

The very fact that, unlike the curve $f_{X|\{(0.749129,0.609756)\}}(x)$*, the right branch of the curve* $f_X(x)$ *shoots infinitely upwards, is purely due to the singularity of the curve* $f_X(x)$ *at the extreme point* $x = 1$*.*

Thus, we conclude,

- *Because of the very fact that the equivalence of* $\mu_2 = 0.609756$*, namely* $\sigma^2 = 0.0485617$ *(* $= \mu_2 - \mu_1^2 = 0.609756 - 0.749129^2$*) lies in between* $\sigma^2_{X,U} = 0.0352582$ *and* $\sigma^2_{X,L} = 0.0596767$ *i.e.*
  $\sigma^2_{X,U} = 0.0352582 < \sigma^2 = 0.0485617 < \sigma^2_{X,L} = 0.0596767$, $f_{X|\{(0.749129,0.609756)\}}(x)$ *preserves the monotonic character of the probability distribution in this case.*

- *Moreover, the differences in the values of the distribution functions with regard to the two above probability distributions, namely the differences* $F_{Beta}(x) - F_X(x)$ *for different values of* $x \in (0, 1)$*, such that* $F_{Beta}(x) = \int_0^x f_X(t)dt$ *and* $F_X(x) = \int_0^x f_{X|\{(0.749129,0.609756)\}}(t)dt$*, are plotted simultaneously as follows:*



Representation of the difference of $F_{X|\{(0.749129,0.609756)\}}(x)$ and $F_X(x)$

which evidently shows that the aforesaid maximum difference is roughly 4% and therefore $f_{X|\{(0.749129,0.609756)\}}(x)$ *approximates* $f_X(x)$ *well enough.*

- *Lastly, the entropy of the probability distribution given by the density* $f_{X|\{(0.749129,0.609756)\}}$ *is given as*

$$-(-2.2750656018479605 + 3.5834317748920568\mu_1 - 0.006347275972631056\mu_2) = -0.405517$$

Example 8.1.10 (A beta distribution with $k_1 > 1$ and $k_2 > 1$). *With subject to $k_1 = 8.27$ and $k_2 = 9.75$, the beta distribution as a uni-extremal bell-shaped probability density function of the random variable X is given by*

$$f_X(x) = 208136(1-x)^{8.75}x^{7.27}, \quad 0 \leq x \leq 1 \tag{8.22}$$

and the computed first two moments of the probability distribution are $\mu_1 = 0.458935$ and $\mu_2 = 0.223676$ respectively.

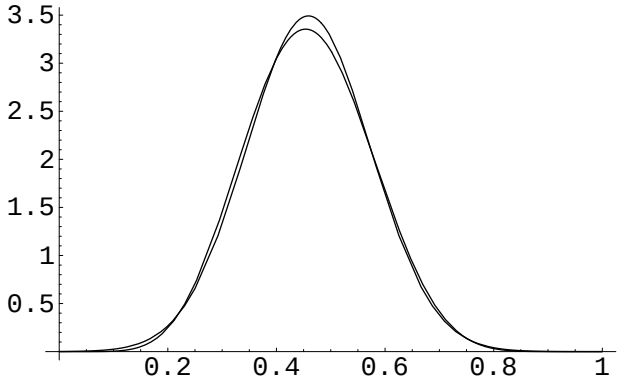
As a matter of fact, due to the bell- shapeliness of the density curve, it is an uni- extremal probability distribution and in fact an uni- modal probability distribution.

Thus, the construction of the fitting appropriate minimum information probability distribution understandably necessitates the availability of at least two moments.

Keeping this in mind, the probability density function of X, viz. $f_{X|\{(0.458935,0.223676)\}}(x)$ giving the computed minimum information probability distribution with subject to $d = (0.458935, 0.223676)$ is given as

$$\begin{aligned} &f_{X|\{(0.458935,0.223676)\}}(x) \\ &= e^{-6.816501847244142+35.15486400876463x-38.30048264870258x^2}, 0 \leq x \leq 1 \end{aligned} \tag{8.23}$$

Clearly, $f_{X|\{(0.458935,0.223676)\}}(x)$ is undoubtedly a very good approximation of $f_X(x)$, which can be easily shown by plotting both of them simultaneously as

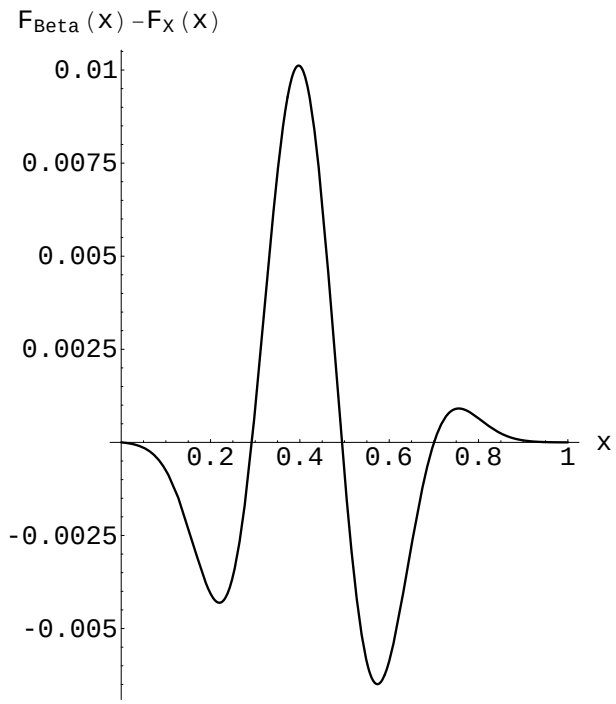


Simultaneous representation of $f_{X|\{(0.458935,0.223676)\}}(x)$ and $f_X(x)$

Thus, we conclude,

- *Because of the very fact that the equivalence of* $\mu_2 = 0.223676$*, namely* $\sigma^2 = 0.0130478$ *(* $= \mu_2 - \mu_1^2 = 0.223676 - 0.458935^2$*) lies well below* $\sigma^2_{X,U} = 0.0795451$ *i.e.*
  $\sigma^2 = 0.0130478 < \sigma^2_{X,U} = 0.0795451$*,* $f_{X|\{(0.458935,0.223676)\}}(x)$ *preserves the uni- modal character of the probability distribution in this case.*

- *Moreover, the differences in the values of the distribution functions with regard to the two above probability distributions, namely the differences* $F_{Beta}(x) - F_X(x)$ *for different values of* $x \in (0,1)$*, such that* $F_{Beta}(x) = \int_0^x f_X(t)dt$ *and* $F_X(x) = \int_0^x f_{X|\{(0.458935,0.223676)\}}(t)dt$*, are plotted simultaneously as follows:*



Representation of the difference of $F_{X|\{(0.458935,0.223676)\}}(x)$ and $F_X(x)$

which evidently shows that the aforesaid maximum difference is roughly 1% and therefore $f_{X|\{(0.458935,0.223676)\}}(x)$ *approximates* $f_X(x)$ *well enough.*

- *Lastly, the entropy of the probability distribution given by the density* $f_{X|\{(0.458935,0.223676)\}}$ *is given as*

$$-(-6.816501847244142 + 35.15486400876463\mu_1 - 38.30048264870258\mu_2) = -0.750661$$

8.1.3 Conclusive remarks

- If the probability distribution of a random variable is not known, the corresponding minimum information probability distribution can be used for the given purpose. For this, the **knowledge of the desired nature** of the probability distribution (for eg. if the probability distribution is monotonic or uni-extremal, etc.) of the random variable under consideration is **of utmost importance**.

  This minimum information probability distribution can be constructed by means of maximum entropy principle with subject to the available moments of the random variable and the goodness of fit of these constructed maximum entropy probability distributions (as we have already illustrated by examples) are remarkably high.

  In order to summarize the essential points with regard to the construction of minimum information probability distributions, we have the following observations from the illustrated examples:

  - If the probability distribution is known to be **monotonic** before hand, we can conclude, that, apart from the range of variability of the random variable, the availability of the first moment is of absolute necessity.
  - If the probability distribution is known to be **uni-extremal** before hand, then, apart from the range of variability of the random variable, the availability of the first two moments is of absolute necessity.

- It is absolutely clear (from the theoretical point of view), that the availability of the moments of higher orders ensures the improvement of the probability distribution of the random variable, provided the additionally introduced moments are contained within certain bounds.

- As far as the **example 8.1.5** of the continuous case is concerned, we know from our algebraic knowledge that the necessary and sufficient requirement for fitting a parabolic curve is the knowledge of two parameters. This explains, why the fitting of our parabolic density curve is almost perfect on usage of the first two moments of the random variable.

This degree of perfectness of the fitting of the same leads us to make a further examination with fitting a non-parabolic density curve as in the **example 8.1.6** of the continuous case, which too resulted an extremely good fitting of the maximum entropy distribution.

- In both discrete and continuous cases, the results show that the fitting of maximum entropy distribution gives excellently good approximations of the original probability distributions.

- However, in case of approximating the beta distribution with the help of the maximum entropy distribution, we are not allowed to ignore the singularities at the end points of the probability density function of the Beta distribution, namely $x = 0$ and $x = 1$. Because of these singularities, the fitting of maximum entropy distribution for two moments could be a big problem, if μ_1 is chosen closer to 0 or 1. In such cases, the story of fitting an approximating maximum entropy probability distribution is different. We shall examine this very fact in the subsequent chapter.

Chapter 9

Comparative studies of beta distributions

In this chapter, we shall make a brief comparative study of **beta distributions** and **uni-extremal minimum information probability distributions**.

9.1 The beta distribution

9.1.1 A brief introduction

With subject to the definition of the beta function $B(k_1, k_2) = \frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(k_1+k_2)}$ for positive values of both k_1 and k_2, the beta distribution of the random variable X is defined by its probability density function $f_X^{(B)}(x)$, $x \in [0, 1]$ as

$$f_X^{(B)}(x) = \frac{\Gamma(k_1 + k_2)}{\Gamma(k_1)\Gamma(k_2)} x^{k_1-1}(1-x)^{k_2-1}, \quad 0 \leq x \leq 1 \tag{9.1}$$

We all know that the continuous beta distribution can be uniform (constant) or monotone (increasing or decreasing) or even an uni-extremal (uni-modal or bathtub shaped) probability distribution. This very fact can be stated as follows for our clarity:

- If $k_1 = k_2 = 1$, then the beta distribution is uniform
- If $k_1 < 1$ and $k_2 \geq 1$, then the beta distribution is strictly monotone decreasing

- If $k_1 \geq 1$ and $k_2 < 1$, then the beta distribution is strictly monotone increasing
- If $k_1 < 1$ and $k_2 < 1$, then the beta distribution is bathtub-shaped
- If $k_1 > 1$ and $k_2 > 1$, then the beta distribution is uni-modal

9.1.2 The distributional moments

Just like an uni-extremal minimum information probability distribution, a beta distribution is also uniquely determinable by its first two moments. We shall establish this very fact in this subsection. However, unlike a monotone minimum information probability distribution, which necessitates the availability of the first moment only, the construction of a beta distribution even for monotone cases necessitates the availability of its first two moments. As usual, let us symbolize the first two moments by μ_1 and μ_2 here.

Now, if the function $f_{(k_1,k_2)} : (0, +\infty)^2 \to (0, 1)^2$ be defined by

$$f_{(k_1,k_2)}(k_1, k_2) = \begin{pmatrix} \frac{k_1}{k_1+k_2} \\ \\ \frac{k_1(k_1+1)}{(k_1+k_2)(k_1+k_2+1)} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \\ \mu_2 \end{pmatrix} \tag{9.2}$$

then with subject to $\mu_1^2 < \mu_2 < \mu_1$, $f_{(k_1,k_2)}$ becomes bijective and is easily invertible to $f^{-1}_{(k_1,k_2)} = f_{(\mu_1,\mu_2)}$, so that $f_{(\mu_1,\mu_2)} : (0, 1)^2 \to (0, +\infty)^2$ is defined by

$$f_{(\mu_1,\mu_2)}(\mu_1, \mu_2) = \begin{pmatrix} \frac{\mu_1-\mu_2}{\mu_2-\mu_1^2}\,\mu_1 \\ \\ \frac{\mu_1-\mu_2}{\mu_2-\mu_1^2}(1-\mu_1) \end{pmatrix} = \begin{pmatrix} k_1 \\ \\ k_2 \end{pmatrix} \tag{9.3}$$

Therefore, we can very well see that for every predetermined pair of moments, namely (μ_1, μ_2), there exists an uniquely determined pair of beta distribution parameters, namely (k_1, k_2). This establishes our assertion.

9.1.3 The entropy

In order to derive the entropy of the beta distribution, we need to define the digamma function at first. The digamma function denoted by $\psi(t)$ is the first derivative of the natural logarithm of the gamma function denoted by $\Gamma(t)$

with respect to the parameter t. This digamma function $\psi(t)$ is therefore given as

$$\psi(t) = \frac{\Gamma^{'}(t)}{\Gamma(t)} = \frac{d}{dt}\left(\log \Gamma(t)\right) \tag{9.4}$$

As the next step, by $\frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(k_1+k_2)} = \int\limits_0^1 x^{k_1-1}(1-x)^{k_2-1}dx$, we can easily see that

$$\frac{\partial}{\partial k_1}\left(\frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(k_1+k_2)}\right) = \int\limits_0^1 x^{k_1-1}(1-x)^{k_2-1}\log x\, dx \tag{9.5}$$

$$\frac{\partial}{\partial k_2}\left(\frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(k_1+k_2)}\right) = \int\limits_0^1 x^{k_1-1}(1-x)^{k_2-1}\log(1-x)\, dx \tag{9.6}$$

Therefore, by using the definition (9.4) on (9.5), we get

$$\begin{aligned}
&\int\limits_0^1 x^{k_1-1}(1-x)^{k_2-1}\log x\, dx \\
&= \frac{\Gamma(k_1+k_2)\Gamma^{'}(k_1) - \Gamma(k_1)\Gamma^{'}(k_1+k_2)}{\left(\Gamma(k_1+k_2)\right)^2}\Gamma(k_2) \\
&= \frac{\Gamma(k_1+k_2)\Gamma(k_1)\psi(k_1) - \Gamma(k_1)\Gamma(k_1+k_2)\psi(k_1+k_2)}{\left(\Gamma(k_1+k_2)\right)^2}\Gamma(k_2) \\
&= \frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(k_1+k_2)}\left(\psi(k_1) - \psi(k_1+k_2)\right) \\
&= B(k_1,k_2)\left(\psi(k_1) - \psi(k_1+k_2)\right)
\end{aligned} \tag{9.7}$$

Exactly in the same way, by using the definition (9.4) on (9.6), we get

$$\int\limits_0^1 x^{k_1-1}(1-x)^{k_2-1}\log(1-x)\, dx = B(k_1,k_2)\left(\psi(k_2) - \psi(k_1+k_2)\right) \tag{9.8}$$

Hence, by using the deductions (9.7) and (9.8), we arrive at the expression of the entropy of the beta distribution with subject to the natural logarithm

without any serious loss of generality, denoted by E_{Beta} as

$$\begin{aligned}
E_{Beta} &= \int_0^1 \frac{x^{k_1-1}(1-x)^{k_2-1}}{B(k_1,k_2)} \log\left(\frac{B(k_1,k_2)}{x^{k_1-1}(1-x)^{k_2-1}}\right) dx \\
&= \log(B(k_1,k_2)) - \frac{k_1-1}{B(k_1,k_2)} \int_0^1 x^{k_1-1}(1-x)^{k_2-1} \log x \, dx \\
&\qquad - \frac{k_2-1}{B(k_1,k_2)} \int_0^1 x^{k_1-1}(1-x)^{k_2-1} \log(1-x) \, dx \qquad (9.9) \\
&= \log(B(k_1,k_2)) - (k_1-1)\Big(\psi(k_1) - \psi(k_1+k_2)\Big) \\
&\qquad - (k_2-1)\Big(\psi(k_2) - \psi(k_1+k_2)\Big) \\
&= \log(B(k_1,k_2)) - (k_1-1)\psi(k_1) - (k_2-1)\psi(k_2) \\
&\qquad + (k_1+k_2-2)\psi(k_1+k_2)
\end{aligned}$$

Note: In the language of Calculus, the natural logarithm is often denoted by log and the common logarithm is often denoted by $\log_{10}$. Therefore, there should not be any confusion in this regard.

9.2 Generalities

Let $F_X(x)$, $0 \leq x \leq 1$ denote the distribution function of the random variable X (X may follow either a beta distribution or a minimum information distribution). With subject to the transformation $Z = 1 - X$, let $F_Z(z)$, $0 \leq z \leq 1$ denote the distribution function of the random variable Z.

Therefore, if $f_X(x)$ and $f_Z(z)$ denote probability densities of X and Z respectively, then the probability differentials of both X and Z, namely $dF_X(x)$ and $dF_Z(z)$, must be the same in terms of magnitude, i.e.

$$|dF_X(x)| = |dF_Z(z)|$$

which gives

$$\begin{aligned} |f_X(x)dx| &= |f_Z(z)dz| \\ \Leftrightarrow f_X(x) &= f_Z(z)\left|\frac{dz}{dx}\right| \\ \Leftrightarrow f_X(x) &= f_Z(z) = f_Z(1-x) \end{aligned} \tag{9.10}$$

Obviously, it is clear that, if $E[X] = \mu_1$ and $Var[X] = \sigma^2$ then $E[Z] = 1 - \mu_1$ and $Var[Z] = \sigma^2$.

Before we proceed, we need to state importantly that we need to denote the probability densities particularly of the beta- and the minimum information probability distributions explicitly. They are denoted by $f^{Beta}_{X|\{d\}}(x)$ and $f^{MEP}_{X|\{d\}}(x)$ respectively for $0 \leq x \leq 1$, where $d = (\mu_1, \mu_2)$ specifies the moments of the probability distribution individually. In the same way, $F^{Beta}_{X|\{d\}}(x)$ and $F^{MEP}_{X|\{d\}}$ denote the distribution functions of the two probability distributions respectively.

However, in general (i.e without any consideration of whether the probability distribution is beta or minimum information), the notations $f_X(x)$ and $F_X(x)$ stand for the probability density function and the distribution function respectively.

Thus, by keeping these things in mind, we proceed to perform our comparative studies. It has to be unforgettably stated that **we shall need to use our software programm referred to the continuous uni-extremal cases for our course of comparative studies**. Apart from this software

program, we also need to use the program *Mathematica* for the construction of graphics.

Our comparative studies shall be in terms of graphical representations, where in each such graphical figure, different statistical sizes (i.e. parameters) are plotted against different values of σ^2 within the permissible range given by $0 < \sigma^2 < \mu_1(1-\mu_1)$, such that μ_1 is fixedly chosen each time in form of $\mu_1 \in \{0.01, 0.05, 0.10, 0.15, 0.20, \ldots, 0.95, 0.99\}$. These graphical figures are individually addressed to

1. the difference of entropies between the two probability distributions for different values of $\sigma^2 = Var[X] = Var[z]$. That is, the difference between the entropy of $f^{MEP}_{X|\{d\}}(x)$ and the same of $f^{Beta}_{X|\{d\}}(x)$ is plotted against σ^2.

2. skewness of both the probability distributions $f^{MEP}_{X|\{d\}}(x)$ and $f^{Beta}_{X|\{d\}}(x)$ are simultaneously plotted against σ^2.

3. left and right kurtosis of both the probability distributions. For the probability density $f_X(x)$, they are defined by $\int_0^{\mu_1} (\frac{x-\mu_1}{\sigma})^4 f_X(x)dx$ and $\int_{\mu_1}^{1} (\frac{x-\mu_1}{\sigma})^4 f_X(x)dx$ respectively. Each of the left and the right kurtosis of both the probability distributions $f^{MEP}_{X|\{d\}}(x)$ and $f^{Beta}_{X|\{d\}}(x)$ are simultaneously plotted against σ^2.

4. maximum difference between the distribution functions of the two probability distributions is plotted against σ^2.

5. minimum difference between the distribution functions of the two probability distributions is plotted against σ^2.

9.3 The entropy

At first we shall show that the entropies of the probability distributions of X and Z are equal. For this, with subject to (9.10) we simply get

$$\begin{aligned}&\int_0^1 f_X(x)\log\left(\frac{1}{f_X(x)}\right)dx = \int_0^1 f_Z(1-x)\log\left(\frac{1}{f_Z(1-x)}\right)dx\\ &= \int_{z=1}^{z=0} f_Z(z)\log\left(\frac{1}{f_Z(z)}\right)d(1-z) = \int_0^1 f_Z(z)\log\left(\frac{1}{f_Z(z)}\right)dz\end{aligned} \tag{9.11}$$

Therefore, it is conclusively clear that, if the entropies of the probability distributions of X and Z with respective probability densities f_X and f_Z be denoted by $Entropy(f_X)$ and $Entropy(f_Z)$ respectively, then we have $Entropy(f_X) = Entropy(f_Z)$.

Importantly, we need to repeat that we must have $Entropy(f_X) < 0$, each time when the probability density f_X is different from 1 (i.e probability density of the constant probability distribution). Moreover, as we already know, $Entropy(f^{MEP}_{X|\{(\mu_1,\sigma^2+\mu_1^2)\}})$ is obviously larger than $Entropy(f^{Beta}_{X|\{(\mu_1,\sigma^2+\mu_1^2)\}})$.

Therefore, the difference $Entropy(f^{MEP}_{X|\{(\mu_1,\sigma^2+\mu_1^2)\}}) - Entropy(f^{Beta}_{X|\{(\mu_1,\sigma^2+\mu_1^2)\}})$, which is positive, is plotted against σ^2 for a fixed value of μ_1 (or $1-\mu_1$) each time.

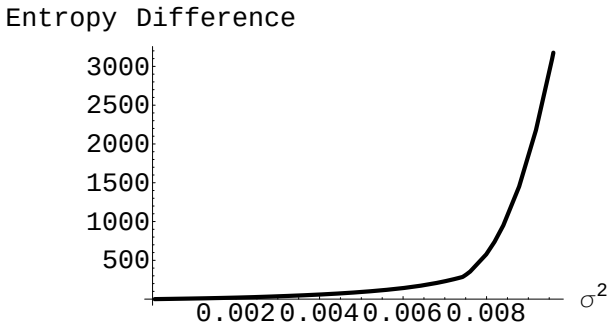


Figure 9.1: Entropy difference against σ^2 for $\mu_1 = 0.01$ or 0.99

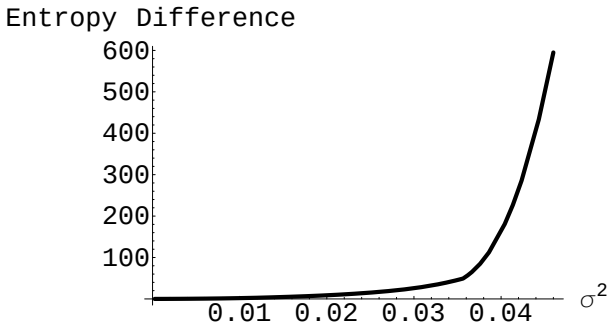


Figure 9.2: Entropy difference against σ^2 for $\mu_1 = 0.05$ or 0.95

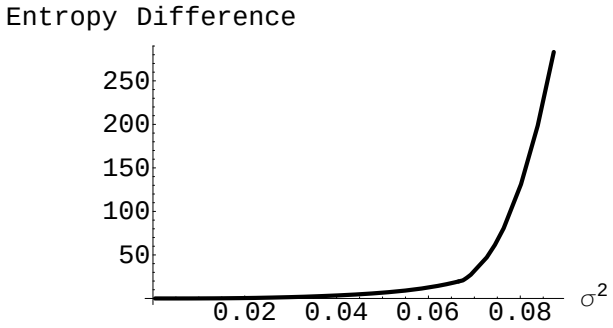


Figure 9.3: Entropy difference against σ^2 for $\mu_1 = 0.10$ or 0.90

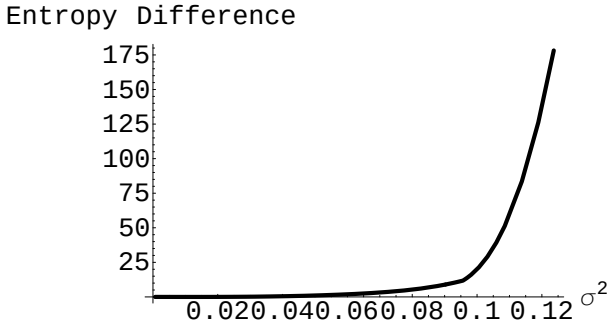


Figure 9.4: Entropy difference against σ^2 for $\mu_1 = 0.15$ or 0.85

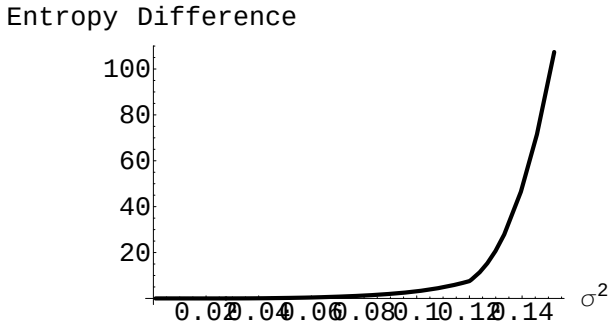


Figure 9.5: Entropy difference against σ^2 for $\mu_1 = 0.20$ or 0.80

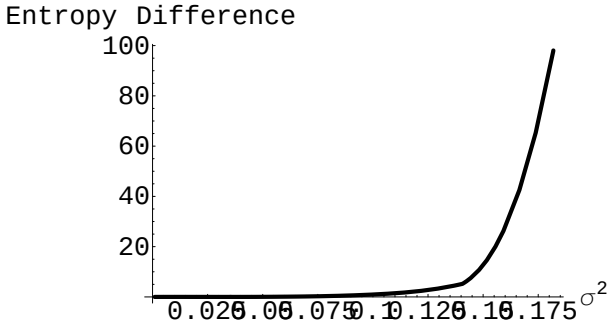


Figure 9.6: Entropy difference against σ^2 for $\mu_1 = 0.25$ or 0.75

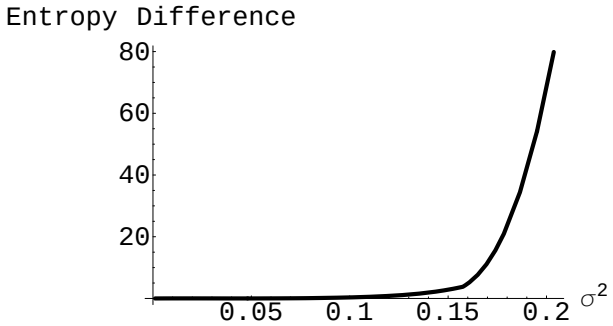


Figure 9.7: Entropy difference against σ^2 for $\mu_1 = 0.30$ or 0.70

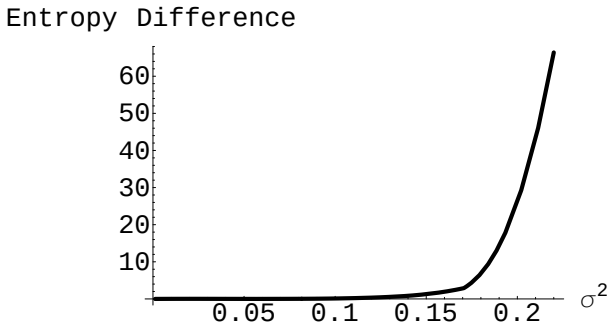


Figure 9.8: Entropy difference against σ^2 for $\mu_1 = 0.35$ or 0.65

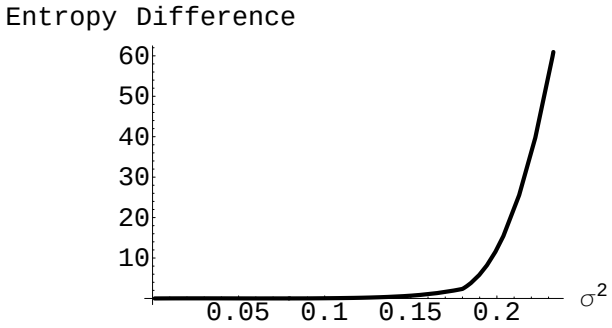


Figure 9.9: Entropy difference against σ^2 for $\mu_1 = 0.40$ or 0.60

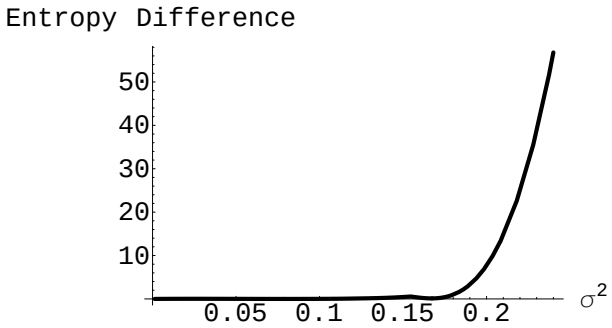


Figure 9.10: Entropy difference against σ^2 for $\mu_1 = 0.45$ or 0.55

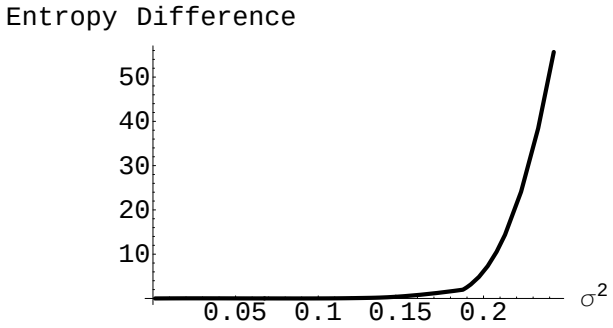


Figure 9.11: Entropy difference against σ^2 for $\mu_1 = 0.50$

Observations:

1. The sharp bend on each of these curves, each time when the curve gets abruptly steeper from a particular value of σ^2, has only to do with the unavoidable technical difficulties in constructing these curves graphically. These curves are basically continuous interpolated curves and these sharp bends are merely the results of interpolating technical difficulties that therefore do not represent any picture of any non-smoothness.

2. The curves for all the considered values of μ_1 have the same characteristics, i.e. the entropy difference rises mildly up to a certain value of σ^2, after which it rises steeply.

3. The curves for values of μ_1 closer to 0.5 show lower values of the entropy difference for higher values of σ^2.

   For eg., for $\mu_1 = 0.45$, the entropy difference for a value of σ^2 closer to $\mu_1(1-\mu_1)$ is about 55, whereas for $\mu_1 = 0.05$, the entropy difference for a value of σ^2 closer to $\mu_1(1-\mu_1)$ is about 600.

   This enables us to conclude that, if μ_1 is allowed to increase from a low value close to 0 to 0.5, then the entropy difference for high values of σ^2 (especially values closer to $\mu_1(1-\mu_1)$) has the tendency to reduce itself.

9.4 Skewness

Let us denote the skewness of the probability distribution of X defined by its probability density f_X by $Sk_{f_X}[X]$. Before we go ahead, we shall at first show that the skewness is negatived by the transformation $X = 1 - Z$, namely $Sk_{f_X}[X] = -Sk_{f_Z}[Z]$ (i.e. the curve turns itself vertically upside down), which can be shown elementarily with the help of (9.10) as

$$\begin{aligned} Sk_{f_X}[X] &= \int_0^1 \left(\frac{x-\mu_1}{\sigma}\right)^3 f_X(x)dx \\ &= -\int_0^1 \left(\frac{1-x-(1-\mu_1)}{\sigma}\right)^3 f_Z(z)dx \\ &= -\int_1^0 \left(\frac{z-(1-\mu_1)}{\sigma}\right)^3 f_Z(z)(-dz) \\ &= \int_0^1 \left(\frac{z-(1-\mu_1)}{\sigma}\right)^3 f_Z(z)dz \\ &= -Sk_{f_Z}[Z] \end{aligned} \tag{9.12}$$

In course of our comparative study, the skewness of each of the two probability distributions are plotted against σ^2 for a fixed μ_1. These two curves are presented simultaneously in the same graphical picture in a way that the plotting of the skewness $Sk_{f_X^{MEP}}[X]$ against σ^2 is **thick-lined** whereas the skewness $Sk_{f_X^{Beta}}[X]$ is **thin-lined**. This makes clear, which curve refers to which probability distribution.

For the sake of simplicity, we shall denote the skewness by ζ_1 in our graphs.

The skewness of both the probability distributions are 0 for $\mu_1 = 0.5$ anyway and therefore the graphical illustration in this case is fruitless and therefore understandably avoided.

With this, we proceed to give the graphical illustrations for different values of μ_1.

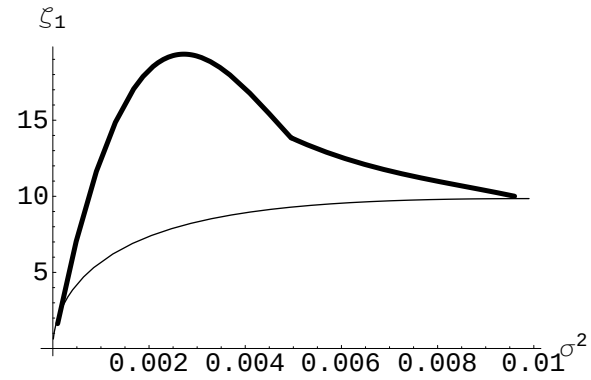


Figure 9.12: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.01$

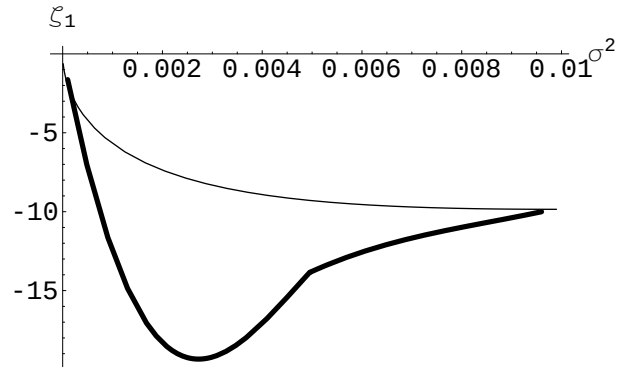


Figure 9.13: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.99$

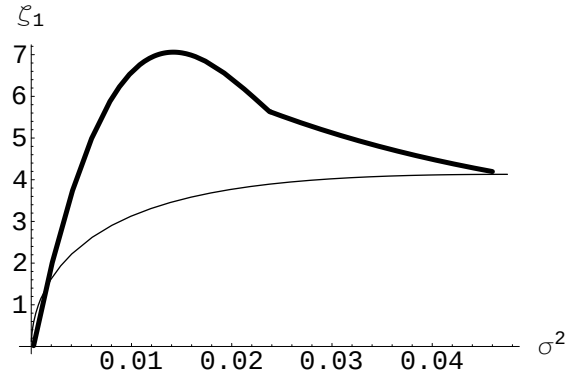


Figure 9.14: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.05$

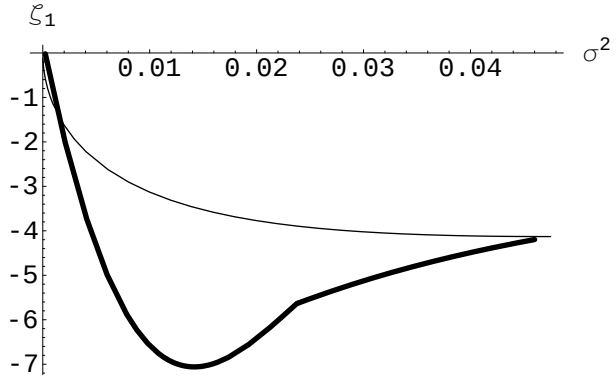


Figure 9.15: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.95$

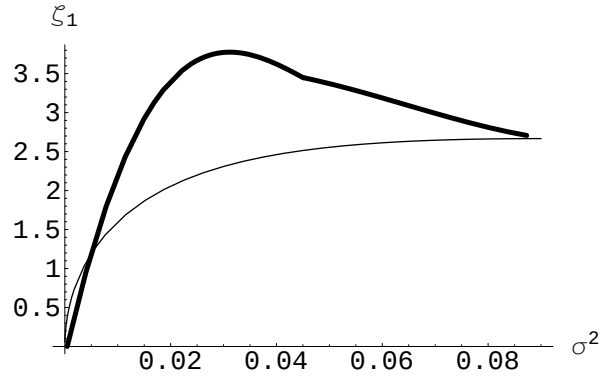


Figure 9.16: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.10$

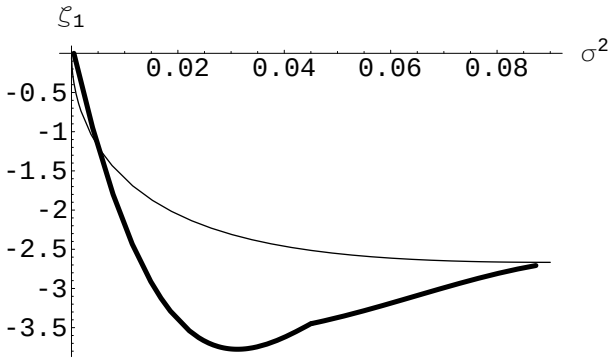


Figure 9.17: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.90$

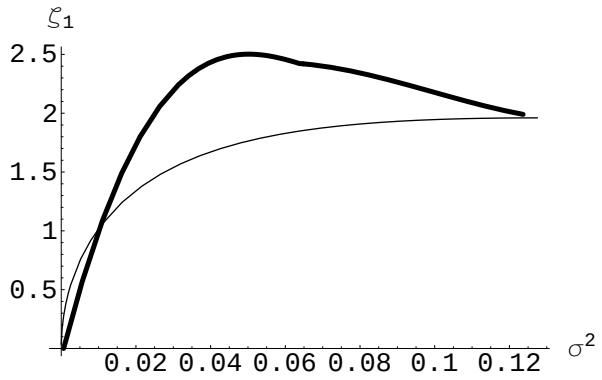


Figure 9.18: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.15$

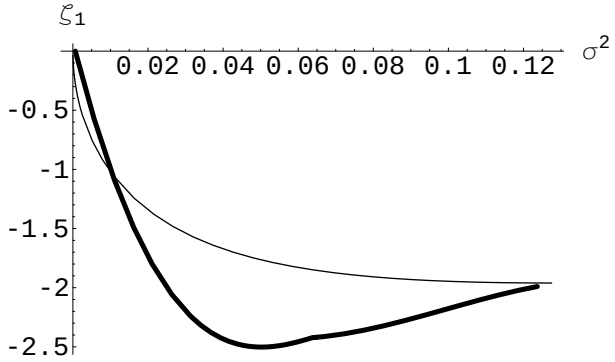


Figure 9.19: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.85$

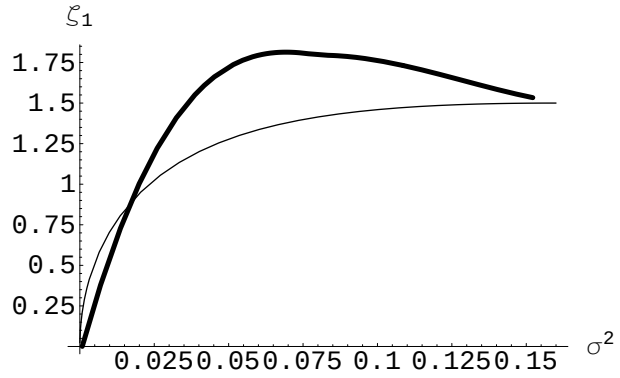


Figure 9.20: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.20$

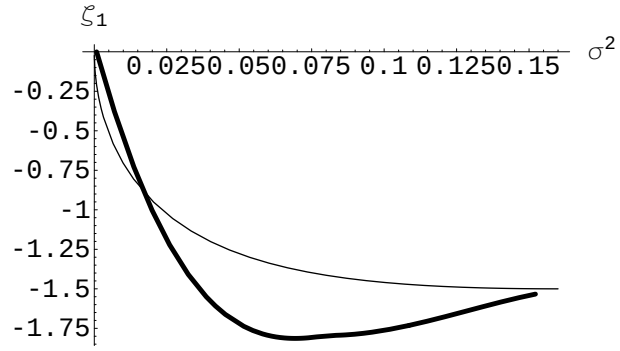


Figure 9.21: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.80$

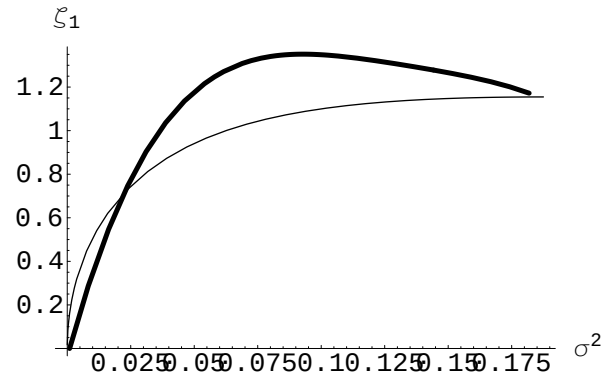


Figure 9.22: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.25$

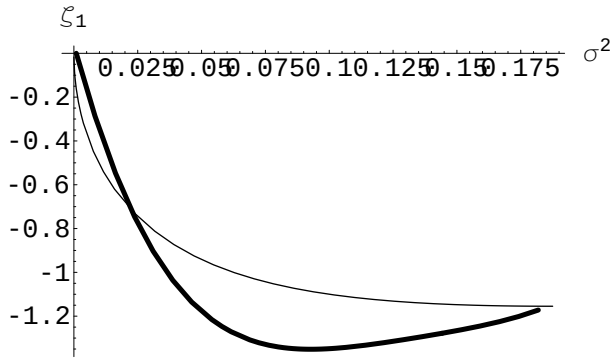


Figure 9.23: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.75$

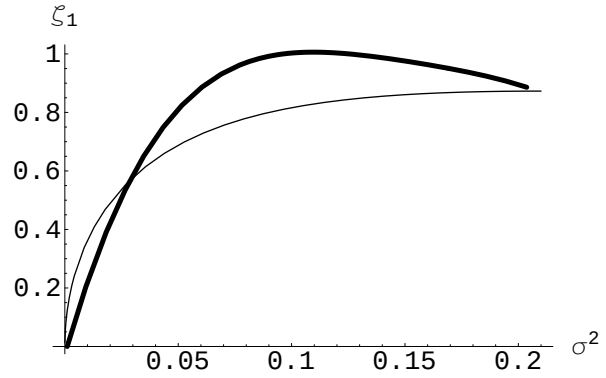


Figure 9.24: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.30$

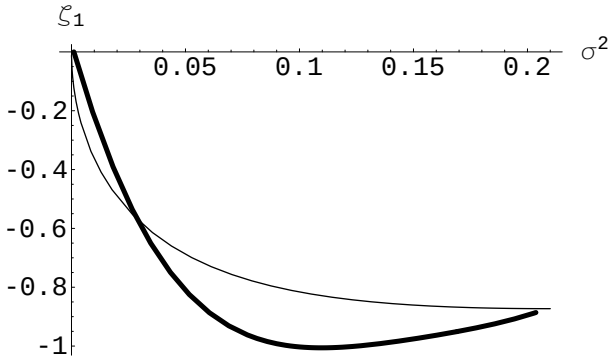


Figure 9.25: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.70$

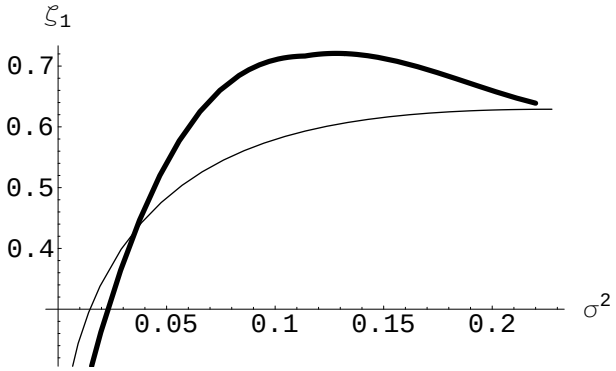


Figure 9.26: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.35$

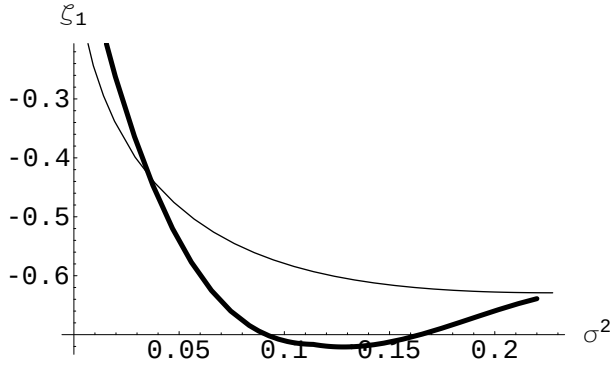


Figure 9.27: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.65$

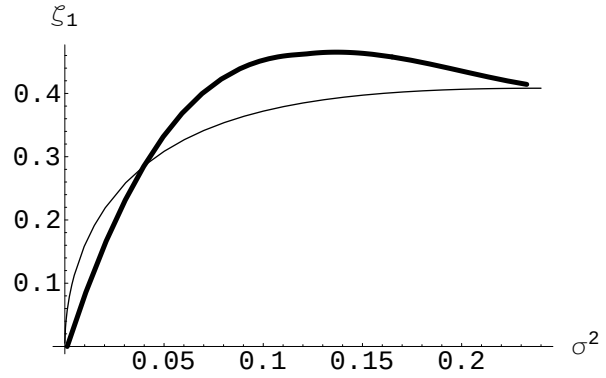


Figure 9.28: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.40$

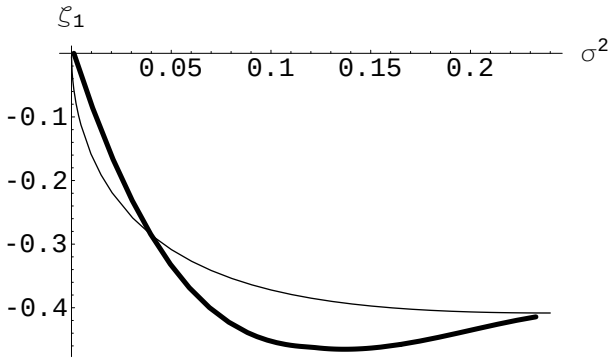


Figure 9.29: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.60$

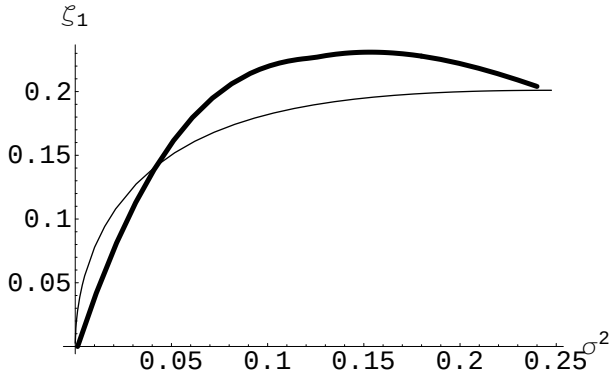


Figure 9.30: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.45$

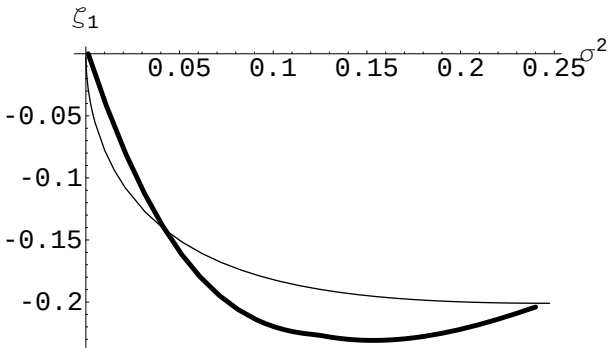


Figure 9.31: Simultaneous curves for the skewness against σ^2 for $\mu_1 = 0.55$

Observations:

1. Even in these cases, the sharp bend in each of the curves is unavoidable because of the said technical difficulties.

   However, this problem is not really difficult in cases for the curves, when μ_1 is taken closer to 0.5.

2. In each of the curves, the skewness of both the probability distributions are almost the same for either low or high values of σ^2. That is, for intermediate values of σ^2, the difference between the skewness of the minimum information distribution and the same of the beta distribution is bit high.

   However, this difference in skewness of the two probability distributions gets gradually lowered as μ_1 is taken closer and closer to 0.5.

3. The curves of both the probability distributions of a given graphical illustration shall understandably come arbitrarily closer to one another, if the fixed value of μ_1 is chosen to be arbitrarily closer to 0.5.

9.5 Kurtosis

Let us denote the left and the right kurtosis of the probability distribution of X defined by its probability density f_X by $Ku_{left,f_X}[X]$ and $Ku_{right,f_X}[X]$ respectively. Before we go ahead, we shall show that $Ku_{left,f_X}[X] = Ku_{right,f_Z}[Z]$ and $Ku_{right,f_X}[X] = Ku_{left,f_Z}[Z]$ as a result of the transformation $X = 1-Z$. This can be shown elementarily with the help of (9.10) as

$$\begin{aligned} Ku_{left,f_X}[X] &= \int_0^{\mu_1} \left(\frac{x-\mu_1}{\sigma}\right)^4 f_X(x)dx \\ &= \int_0^{\mu_1} \left(\frac{1-x-(1-\mu_1)}{\sigma}\right)^4 f_Z(z)dx \\ &= \int_1^{1-\mu_1} \left(\frac{z-(1-\mu_1)}{\sigma}\right)^4 f_Z(z)(-dz) \\ &= \int_{1-\mu_1}^{1} \left(\frac{z-(1-\mu_1)}{\sigma}\right)^4 f_Z(z)dz \\ &= Ku_{right,f_Z}[Z] \end{aligned} \tag{9.13}$$

and

$$\begin{aligned} Ku_{right,f_X}[X] &= \int_{\mu_1}^{1} \left(\frac{x-\mu_1}{\sigma}\right)^4 f_X(x)dx \\ &= \int_{\mu_1}^{1} \left(\frac{1-x-(1-\mu_1)}{\sigma}\right)^4 f_Z(z)dx \\ &= \int_{1-\mu_1}^{0} \left(\frac{z-(1-\mu_1)}{\sigma}\right)^4 f_Z(z)(-dz) \\ &= \int_0^{1-\mu_1} \left(\frac{z-(1-\mu_1)}{\sigma}\right)^4 f_Z(z)dz \\ &= Ku_{left,f_Z}[Z] \end{aligned} \tag{9.14}$$

For the sake of simplicity, we shall denote the left and the right kurtosis by ζ_{21} and ζ_{22} respectively in our graphs.

With this, we proceed to give the graphical illustrations for different values of μ_1.

However, we must unforgettably state a very important thing: In course of plotting the simultaneous curves for the left kurtosis against variance for $\mu_1 = 0.01$ and $\mu_1 = 0.05$, we had to face technical difficulties in plotting the curves for the entire allowable variance, namely $\sigma^2 < \mu_1(1 - \mu_1)$. In those cases, we had restrict the ranges to $\sigma^2 < 0.0003$ and $\sigma^2 < 0.025$ respectively. However, these restrictions did not hamper our comparative study at all.

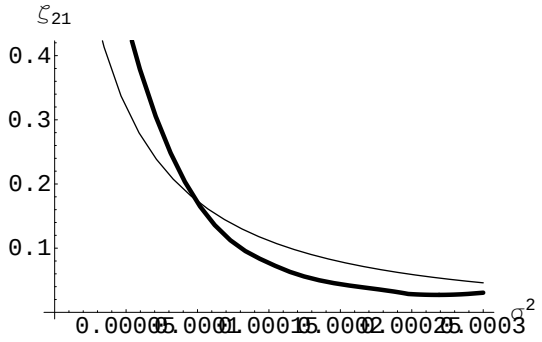


Figure 9.32: Simultaneous curves for the **left** kurtosis against $\sigma^2 \leq 0.0003$ for $\mu_1 = 0.01$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.99$

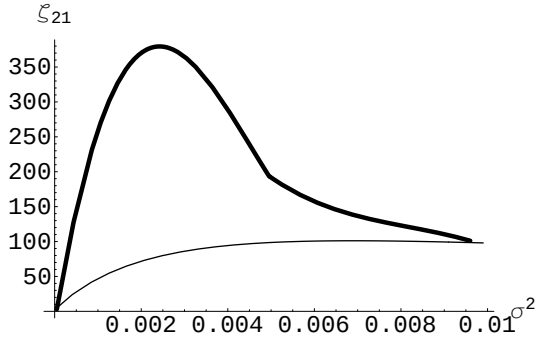


Figure 9.33: Simultaneous curves for the **left** kurtosis against σ^2
for $\mu_1 = 0.99$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.01$

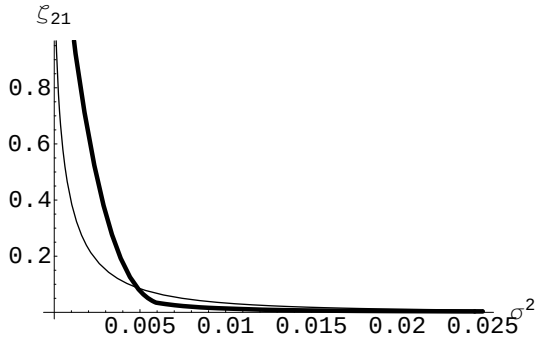


Figure 9.34: Simultaneous curves for the **left** kurtosis against $\sigma^2 \leq 0.025$ for $\mu_1 = 0.05$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.95$

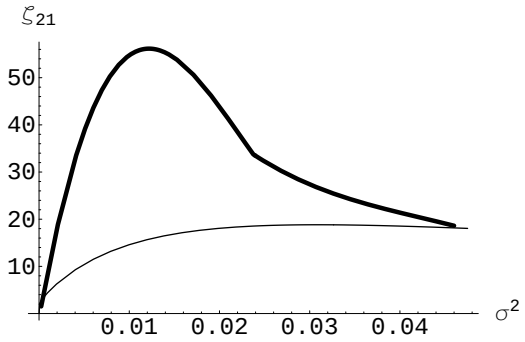


Figure 9.35: Simultaneous curves for the **left** kurtosis against σ^2
for $\mu_1 = 0.95$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.05$

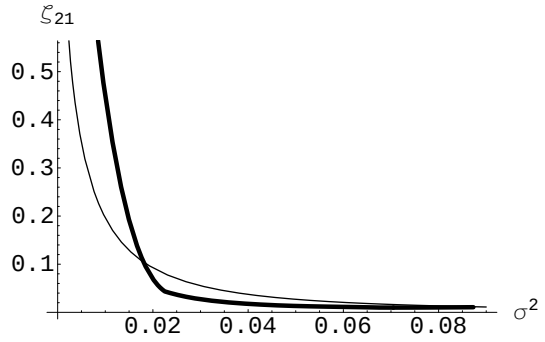


Figure 9.36: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.10$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.90$

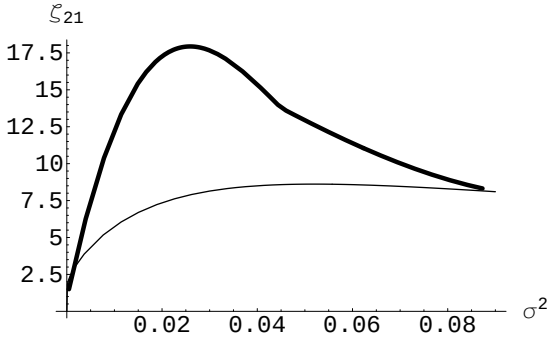


Figure 9.37: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.90$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.10$

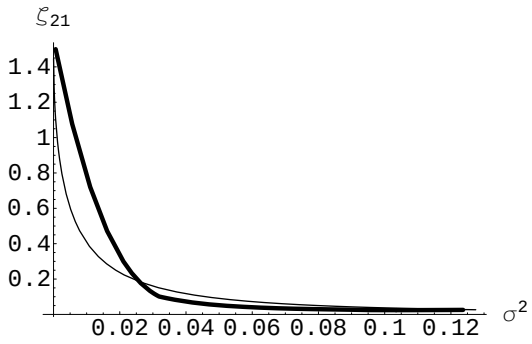


Figure 9.38: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.15$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.85$

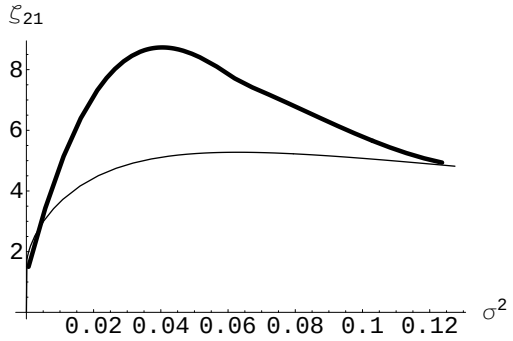


Figure 9.39: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.85$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.15$

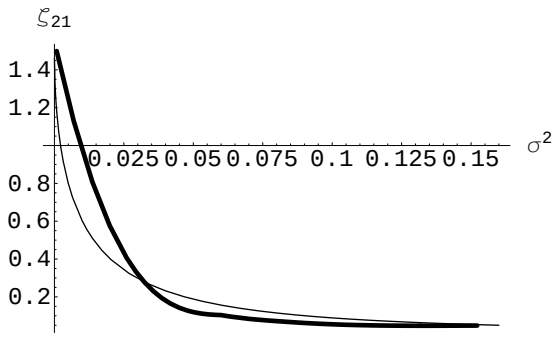


Figure 9.40: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.20$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.80$

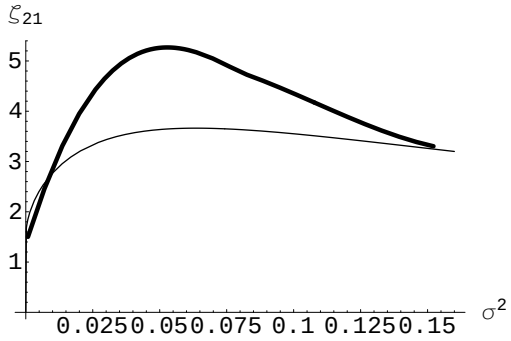


Figure 9.41: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.80$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.20$

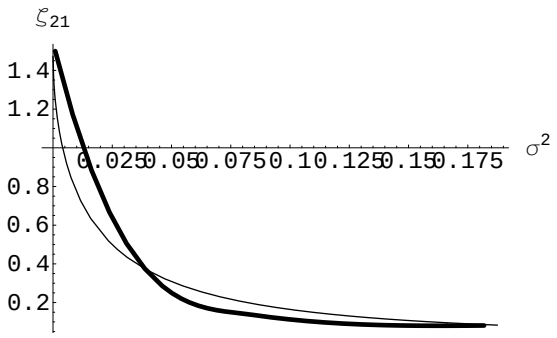


Figure 9.42: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.25$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.75$

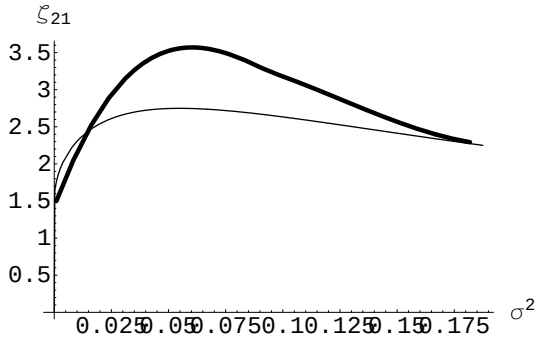


Figure 9.43: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.75$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.25$

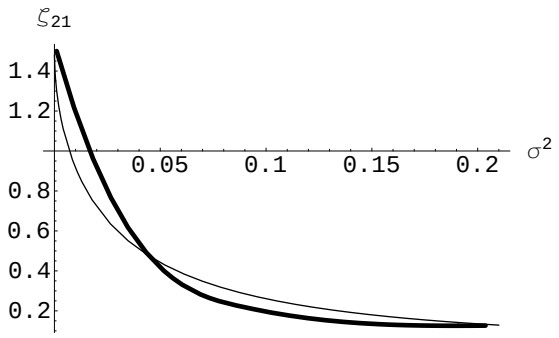


Figure 9.44: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.30$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.70$

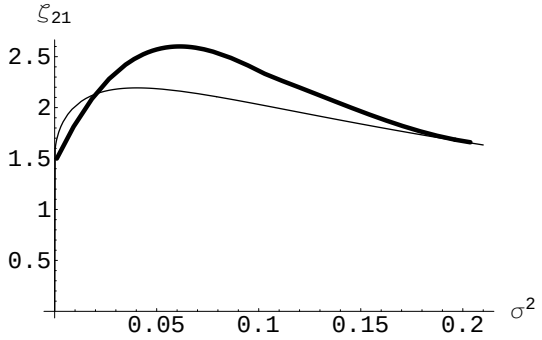


Figure 9.45: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.70$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.30$

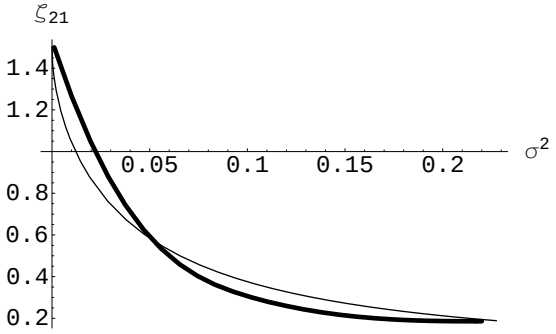


Figure 9.46: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.35$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.65$

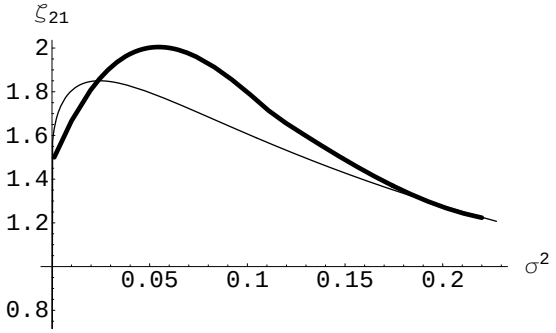


Figure 9.47: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.65$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.35$

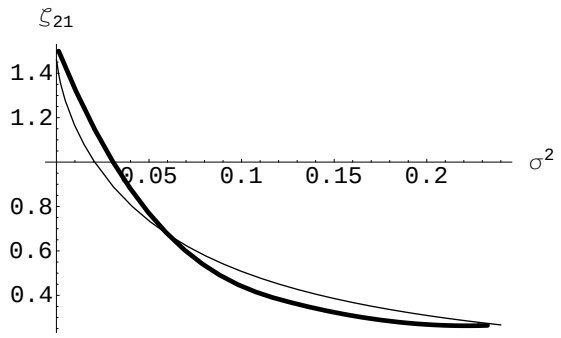


Figure 9.48: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.40$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.60$

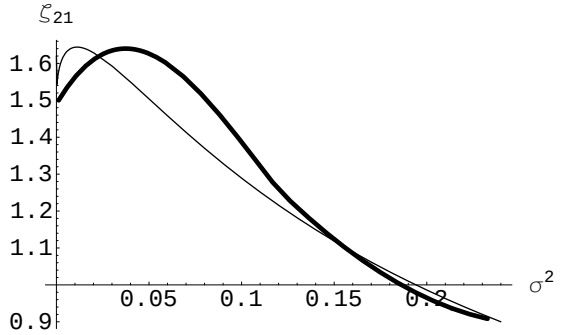


Figure 9.49: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.60$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.40$

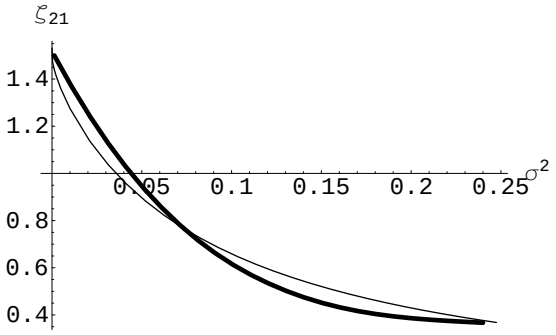


Figure 9.50: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.45$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.55$

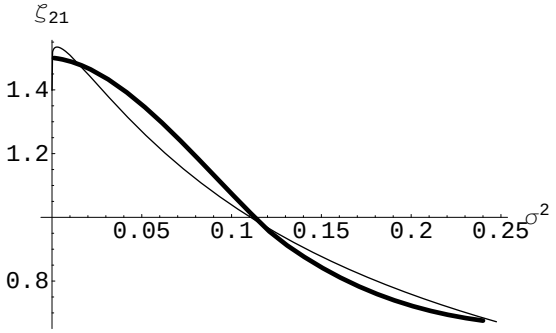


Figure 9.51: Simultaneous curves for the **left** kurtosis against σ^2 for $\mu_1 = 0.55$ **as well as** the same for the **right** kurtosis for $\mu_1 = 0.45$

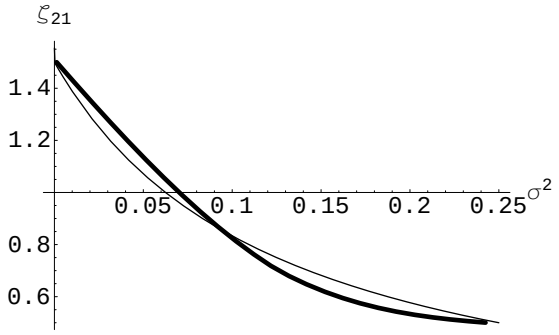


Figure 9.52: Simultaneous curves for the **left** (or **right**) kurtosis against σ^2 for $\mu_1 = 0.50$

Observations:

1. The technical problems giving rise to a sharp bend in a curve did persist. However, for the curves corresponding to the values of μ_1 closer to 0.5 do not have this problem.

2. The difference in the left kurtosis of the probability distributions presented by the curves corresponding to the values of μ_1 significantly lower than 0.5 (or equivalently the difference in the right kurtosis of the probability distributions presented by the curves corresponding to the values of μ_1 significantly higher than 0.5) are high.

   This difference gets gradually lowered as μ_1 is taken closer to 0.5.

3. The kurtosis basically describes the degree of peakedness of a probability distribution. Understandably the increase in variance of a probability distribution reduces the peakedness of the probability distribution. Exactly this phenomenon has been studied here, in fact both the left and the right kurtosis have shown to have got reduced with the increase in variance.

9.6 Maximum and minimum differences

This section shall involve the maximum and the minimum differences between $F^{Beta}_{X|\{d\}}(x)$ and $F^{MEP}_{X|\{d\}}(x)$ for $d = (\mu_1, \sigma^2 + \mu_1^2)$, such that μ_1 is fixed but σ^2 varies within its own range of validity.

The relationship between the maximum and the minimum of these differences necessitates the establishment of a relationship between $F^{Beta}_{X|\{d\}}(x) - F^{MEP}_{X|\{d\}}(x)$ and $F^{Beta}_{Z|\{d_z\}}(z) - F^{MEP}_{Z|\{d_z\}}(z)$, such that $d_z = (1 - \mu_1, \sigma^2 + (1 - \mu_1)^2)$.

Let us see to the following result at first, which necessitates the use of (9.10):

$$\begin{aligned} F_X(x) &= \int_0^x f_X(t)dt = \int_0^x f_Z(1-t)dt \\ &= \int_{u=1}^{u=1-x} f_Z(u)(-du) \\ &= -\int_1^{1-x} f_Z(u)du \\ &= -\left(\int_1^0 f_Z(u)du + \int_0^{1-x} f_Z(u)du\right) \\ &= -\left(-1 + F_Z(1-x)\right) \\ &= 1 - F_Z(z) \end{aligned} \tag{9.15}$$

which brings us to the following relationship:

$$F^{Beta}_{X|\{d\}}(x) - F^{MEP}_{X|\{d\}}(x) = -\left(F^{Beta}_{Z|\{d_z\}}(z) - F^{MEP}_{Z|\{d_z\}}(z)\right) \tag{9.16}$$

and consequently,

$$\max\left(F^{Beta}_{X|\{d\}}(x) - F^{MEP}_{X|\{d\}}(x)\right) = \min\left(F^{Beta}_{Z|\{d_z\}}(z) - F^{MEP}_{Z|\{d_z\}}(z)\right) \tag{9.17}$$

and this necessarily means that the minimum difference curve corresponding to $1 - \mu_1$ is simply the **negative** (i.e. the curve turns itself vertically upside down) of the maximum difference curve corresponding to μ_1.

With this, we proceed to give the graphical illustrations.

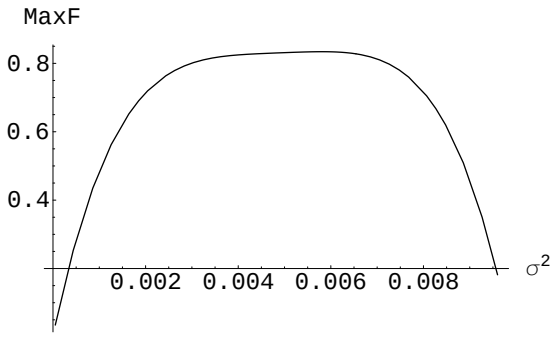


Figure 9.53: **Maximum** difference against σ^2 for $\mu_1 = 0.01$

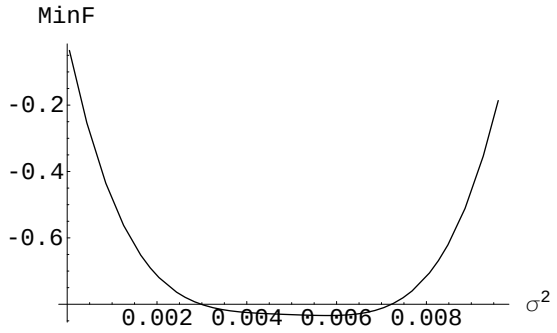


Figure 9.54: **Minimum** difference against σ^2 for $\mu_1 = 0.99$

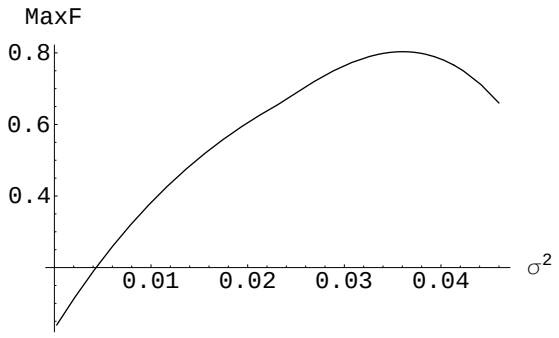


Figure 9.55: **Maximum** difference against σ^2 for $\mu_1 = 0.05$

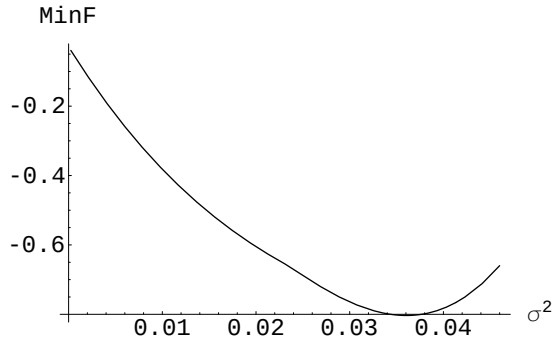


Figure 9.56: **Minimum** difference against σ^2 for $\mu_1 = 0.95$

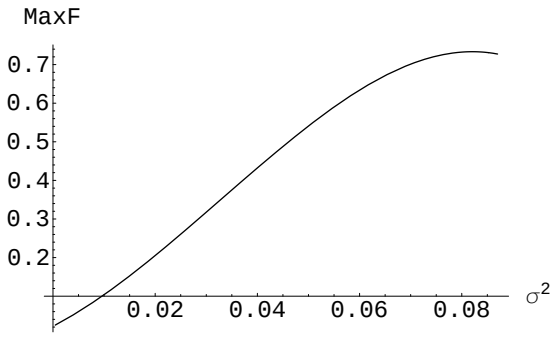


Figure 9.57: **Maximum** difference against σ^2 for $\mu_1 = 0.10$

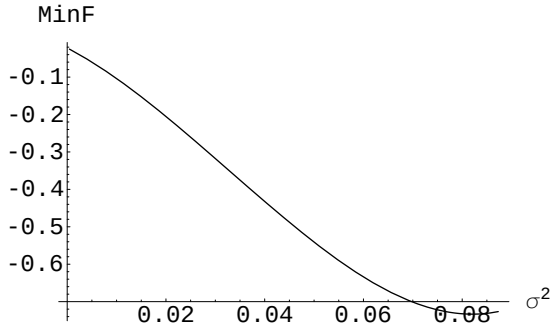


Figure 9.58: **Minimum** difference against σ^2 for $\mu_1 = 0.90$

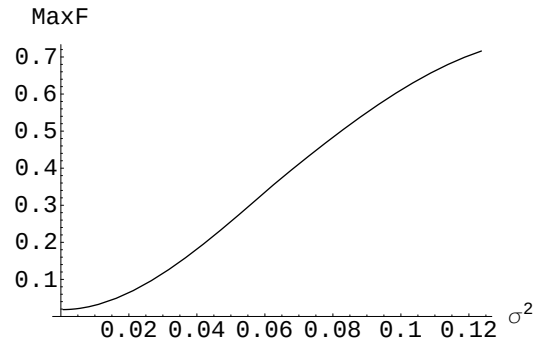


Figure 9.59: **Maximum** difference against σ^2 for $\mu_1 = 0.15$

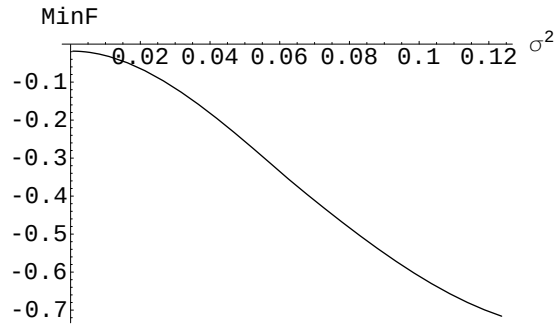


Figure 9.60: **Minimum** difference against σ^2 for $\mu_1 = 0.85$

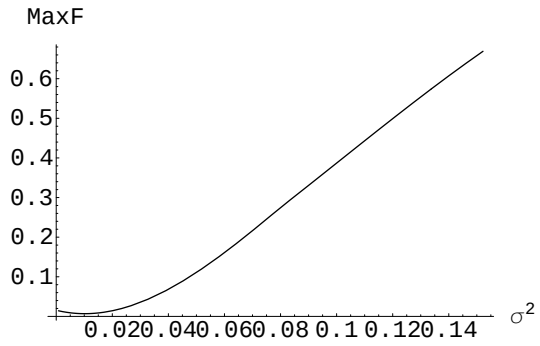


Figure 9.61: **Maximum** difference against σ^2 for $\mu_1 = 0.20$

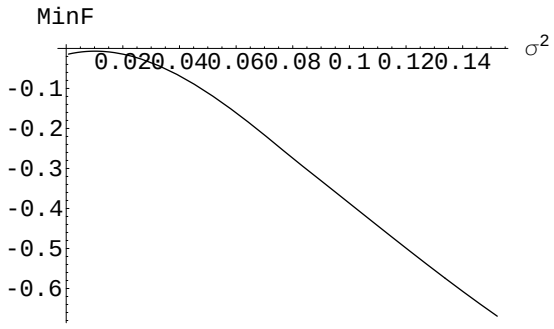


Figure 9.62: **Minimum** difference against σ^2 for $\mu_1 = 0.80$

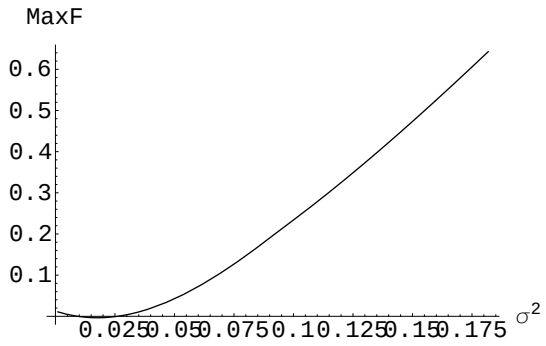


Figure 9.63: **Maximum** difference against σ^2 for $\mu_1 = 0.25$

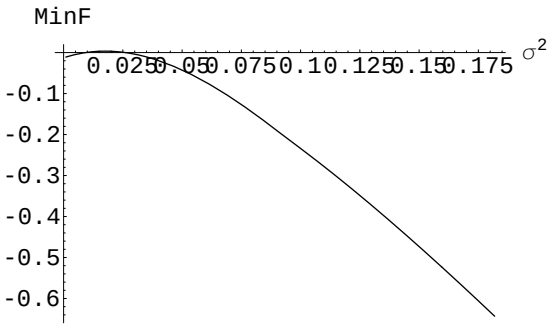


Figure 9.64: **Minimum** difference against σ^2 for $\mu_1 = 0.75$

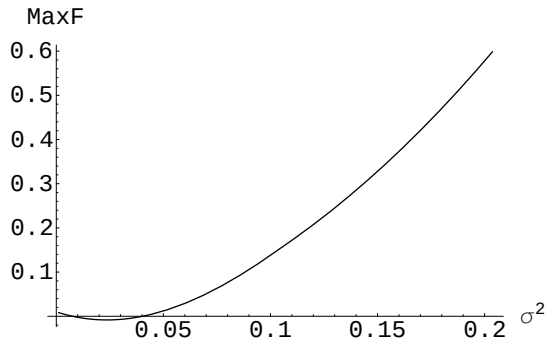


Figure 9.65: **Maximum** difference against σ^2 for $\mu_1 = 0.30$

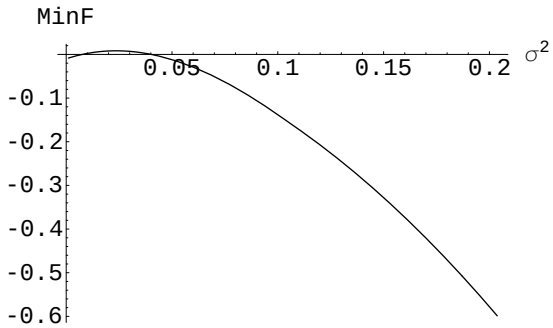


Figure 9.66: **Minimum** difference against σ^2 for $\mu_1 = 0.70$

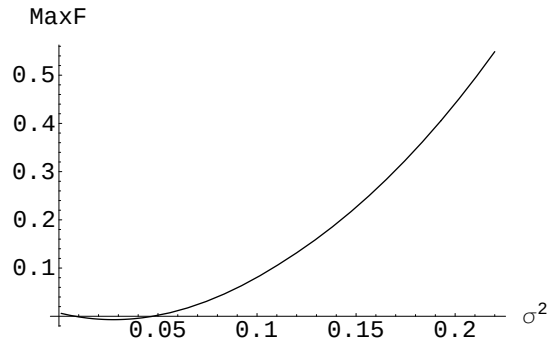


Figure 9.67: **Maximum** difference against σ^2 for $\mu_1 = 0.35$

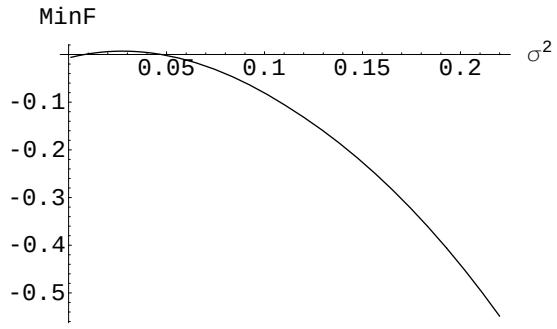


Figure 9.68: **Minimum** difference against σ^2 for $\mu_1 = 0.65$

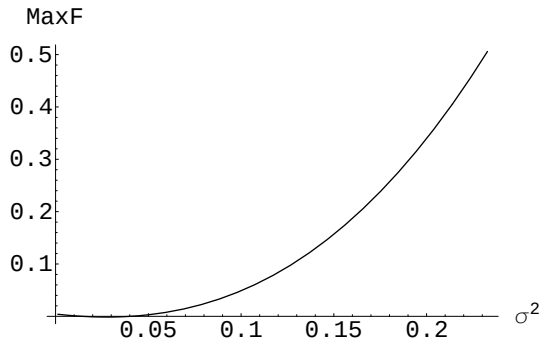


Figure 9.69: **Maximum** difference against σ^2 for $\mu_1 = 0.40$

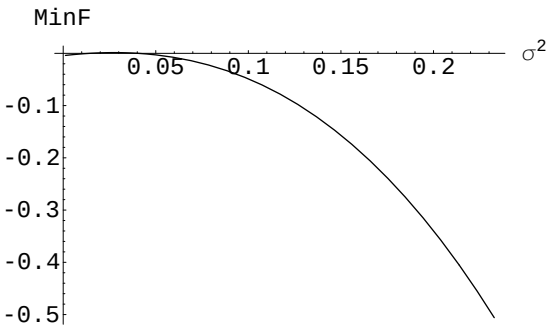


Figure 9.70: **Minimum** difference against σ^2 for $\mu_1 = 0.60$

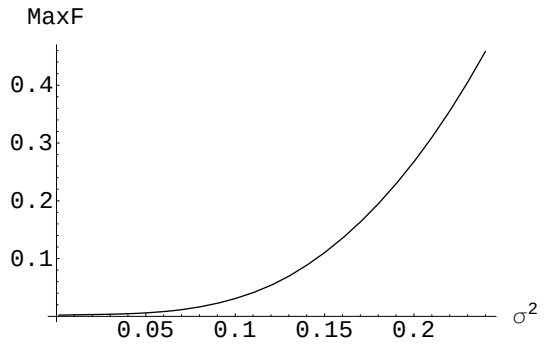


Figure 9.71: **Maximum** difference against σ^2 for $\mu_1 = 0.45$

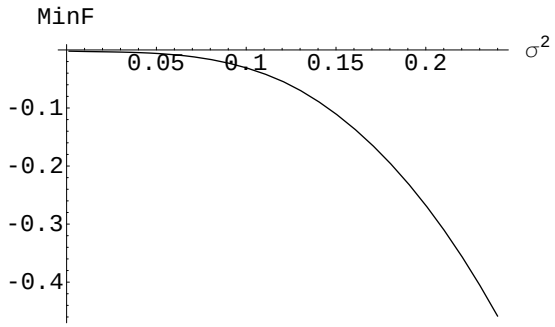


Figure 9.72: **Minimum** difference against σ^2 for $\mu_1 = 0.55$

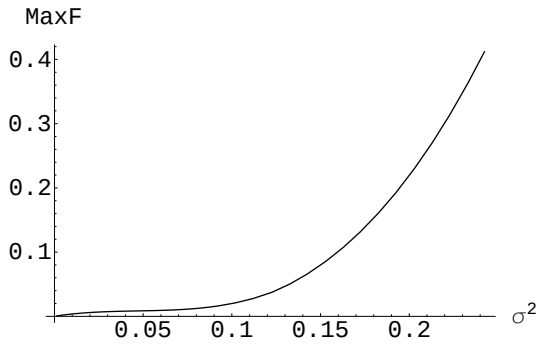


Figure 9.73: **Maximum** difference against σ^2 for $\mu_1 = 0.50$

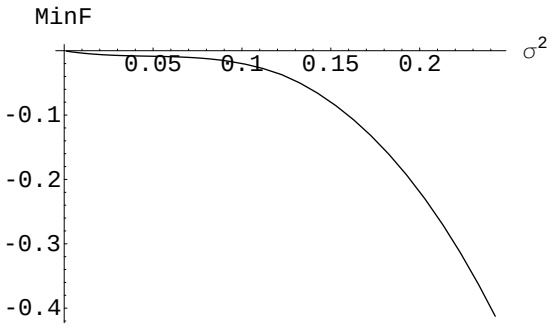


Figure 9.74: **Minimum** difference against σ^2 for $\mu_1 = 0.50$

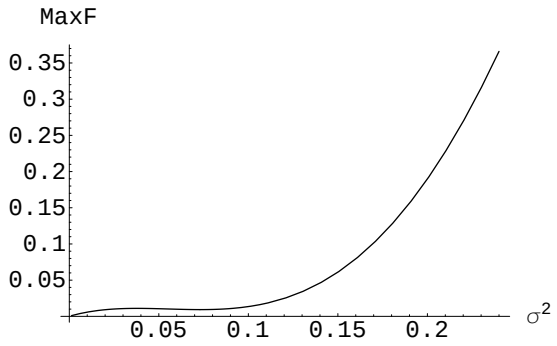


Figure 9.75: **Maximum** difference against σ^2 for $\mu_1 = 0.55$

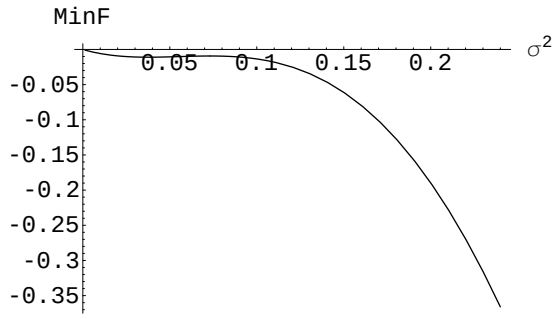


Figure 9.76: **Minimum** difference against σ^2 for $\mu_1 = 0.45$

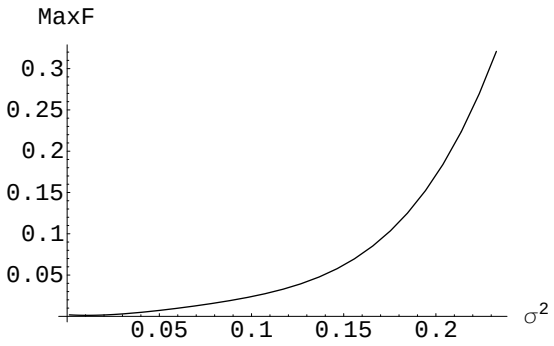


Figure 9.77: **Maximum** difference against σ^2 for $\mu_1 = 0.60$

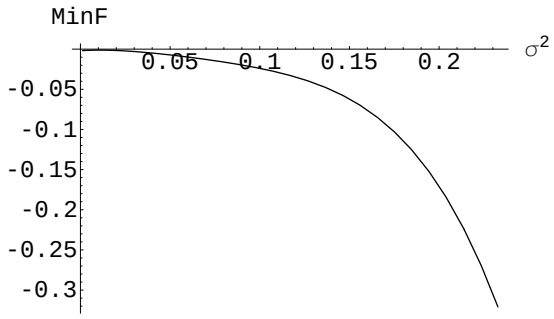


Figure 9.78: **Minimum** difference against σ^2 for $\mu_1 = 0.40$

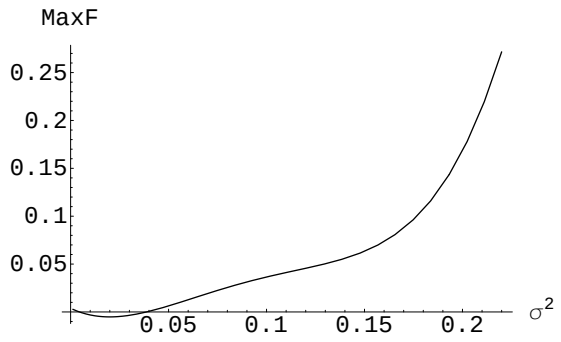


Figure 9.79: **Maximum** difference against σ^2 for $\mu_1 = 0.65$

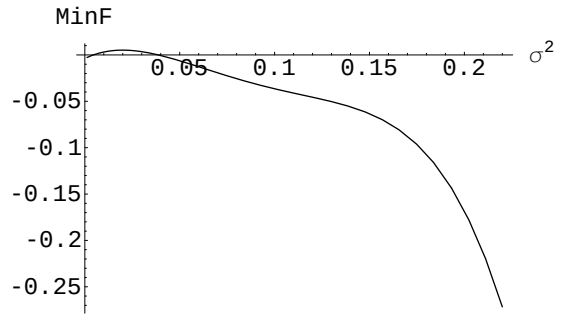


Figure 9.80: **Minimum** difference against σ^2 for $\mu_1 = 0.35$

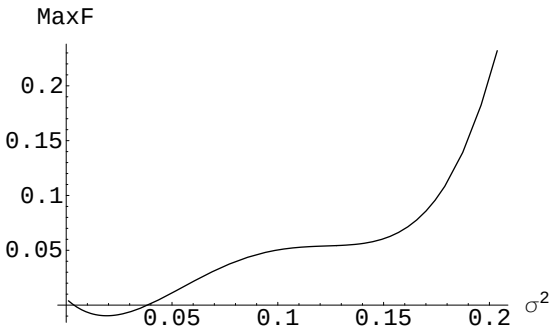


Figure 9.81: **Maximum** difference against σ^2 for $\mu_1 = 0.70$

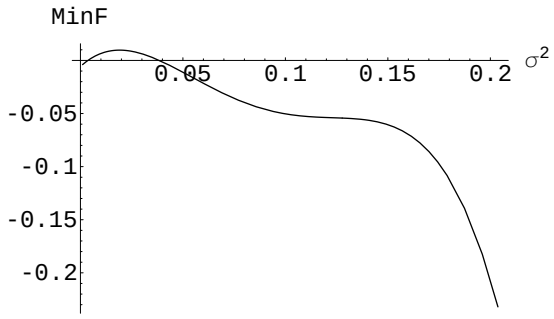


Figure 9.82: **Minimum** difference against σ^2 for $\mu_1 = 0.30$

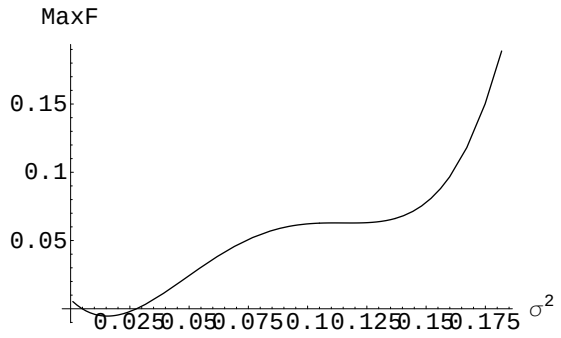


Figure 9.83: **Maximum** difference against σ^2 for $\mu_1 = 0.75$

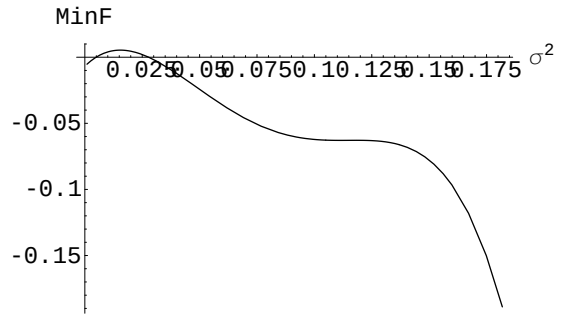


Figure 9.84: **Minimum** difference against σ^2 for $\mu_1 = 0.25$

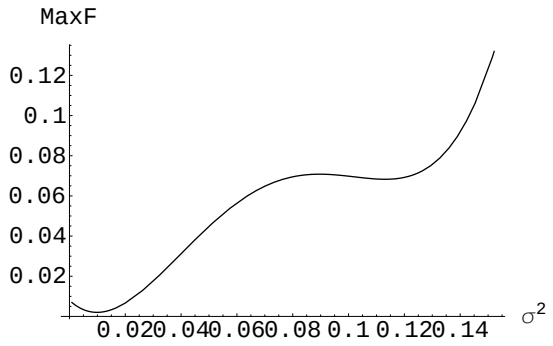


Figure 9.85: **Maximum** difference against σ^2 for $\mu_1 = 0.80$

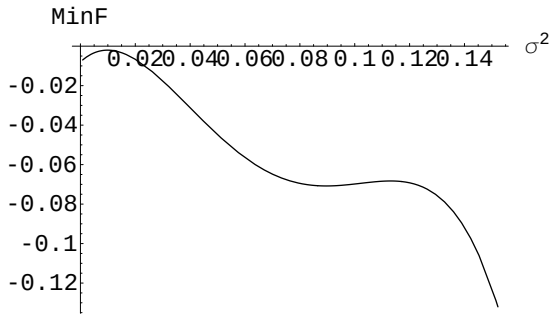


Figure 9.86: **Minimum** difference against σ^2 for $\mu_1 = 0.20$

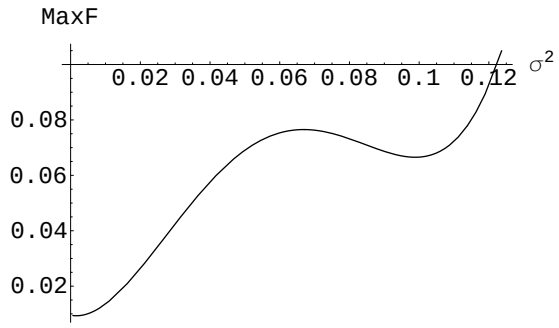


Figure 9.87: **Maximum** difference against σ^2 for $\mu_1 = 0.85$

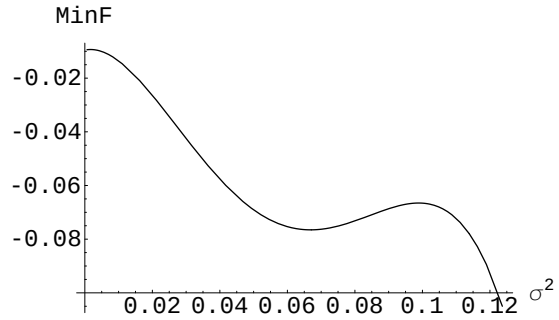


Figure 9.88: **Minimum** difference against σ^2 for $\mu_1 = 0.15$

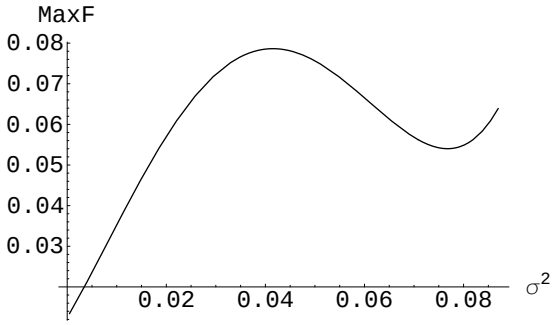


Figure 9.89: **Maximum** difference against σ^2 for $\mu_1 = 0.90$

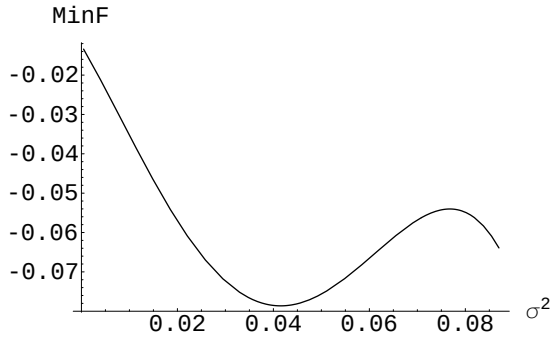


Figure 9.90: **Minimum** difference against σ^2 for $\mu_1 = 0.10$

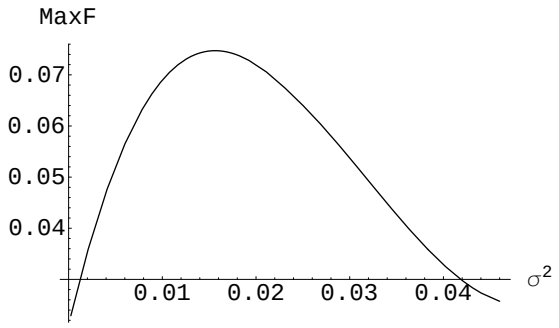


Figure 9.91: **Maximum** difference against σ^2 for $\mu_1 = 0.95$

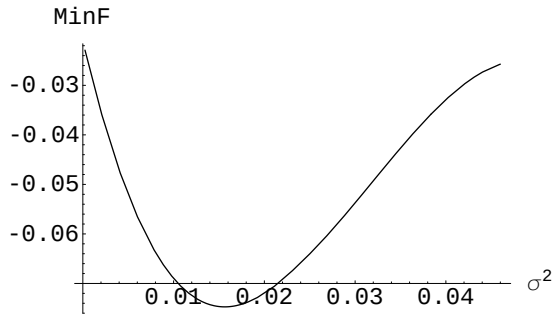


Figure 9.92: **Minimum** difference against σ^2 for $\mu_1 = 0.05$

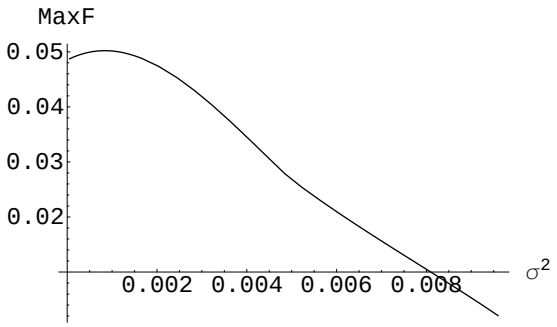


Figure 9.93: **Maximum** difference against σ^2 for $\mu_1 = 0.99$

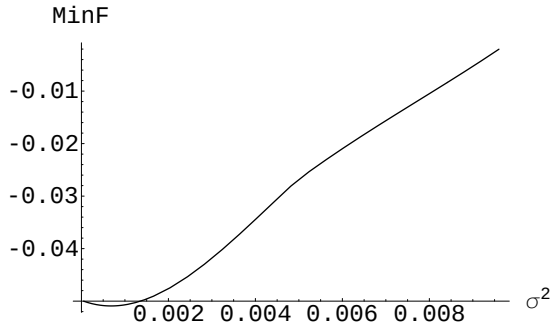


Figure 9.94: **Minimum** difference against σ^2 for $\mu_1 = 0.01$

Observations:

1. Before we go ahead, it must be unforgettably stated that the singularities of the density function of the beta distribution at $x = 0$ and $x = 1$ play a deciding role in this particular comparative study. In fact, for higher values of the chosen variance σ^2, the probabilities of the events given by the intervals $(0, \epsilon_1]$ and $[\epsilon_2, 1)$ are significantly high, even for small values of $\epsilon_1 > 0$ and $\epsilon_2 > 0$.

2. Exactly in this course of studying the behavior of the maximum or minimum differences against variance, the presence of these singularities make a huge difference between the two probability distributions.

3. Since we have already seen that a maximum difference curve corresponding to μ_1 is simply the minimum difference curve corresponding to $1 - \mu_1$ that is simply vertically upside down, **without any loss of generality**, we can confine our discussions to the maximum difference curves only.

4. In general, the maximum values of the maximum differences tend to decrease when μ_1 is chosen to be closer to 1. This can be well explained by the very fact that the problem caused by the singularity at $x = 0$ of the beta distribution density curve becomes more and more insignificant.

   In other words, this very phenomenon can be explained in the following way: By keeping in mind that μ_1 is the position parameter, for a high value of σ^2, the difference $F^{Beta}_{X|\{d\}}(x) - F^{MEP}_{X|\{d\}}(x)$ is larger for a smaller values of x, because the singularity of the density function of the beta distribution at $x = 0$ contributes a higher probability value. On the other hand, because of this singularity, for any particular value of x, the difference $F^{Beta}_{X|\{d\}}(x) - F^{MEP}_{X|\{d\}}(x)$ is larger for smaller values of μ_1.

5. Broadly speaking, there is no real monotonicity in the maximum difference curves, but a lot depends on the value of the chosen μ_1.

6. Now, let us explain the behavior of a maximum difference curve briefly: In general, for any given value of μ_1, if σ^2 is made to increase gradually from its lowest possible presentable value, say 0.0001, we know that the density curve of the beta distribution is expectedly uni-modal. For

this value of σ^2, the maximum difference is understandably very small (i.e. of a small magnitude).

Now, by increasing the σ^2 corresponding to that μ_1, this uni-modal nature slowly fades and slowly turns out to be of monotonic nature. Consequently, the maximum difference increases rapidly because of the aforesaid singularities at $x = 0$ or $x = 1$. This rapidness is even more severe, if $\mu_1 < 0.5$ (i.e. in this case, the monotonic nature is ascribed to the monotonic *decreasing* nature), for which the aforesaid singularity at $x = 0$ plays a predominant role.

By continuing to increase the value of σ^2, a stage is reached, when the maximum difference reaches its maximum value. The reason for this is, the aforesaid monotonic nature discontinues and the density curve of the beta distribution begins to be of bathtub nature. Consequently, because of this bathtub nature, the aforesaid singularity at $x = 1$ starts to play its role and subsequently the maximum difference decreases with the further increase in σ^2. This is because, the probability part generated by the singularity at $x = 1$ increases. This character is particularly visible for $\mu_1 > 0.5$ (i.e. in this case, the monotonic nature, which was present, is ascribed to the monotonic *increasing* nature). This decreasing takes place up to a certain point, after which the maximum difference increases again with the increase in σ^2. This increase is simply explained by the very fact that the probability part generated by the aforesaid singularity at $x = 1$ increases rapidly. This character is again particularly visible for $\mu_1 > 0.5$ cases.

Obviously, this increasing of the maximum difference continues, till the plotting of the curve is discontinued at a point, say $\sigma^2 = \mu_1(1 - \mu_1) - 0.0001$.

General remarks:

1. Unlike a uni-extremal minimum entropy information probability distribution, beta distribution possesses important characteristics because of its singularities at points $x = 0$ and $x = 1$ of its density curves in cases for $k_1 < 0, k_2 < 0$.

2. Because of these singularities, the calculated probabilities even within small neighborhoods of $x = 0$ and $x = 1$ are significantly high, for eg. the beta distribution with parameters $k_1 = 0.03$ and $k_2 = 2.97$, where $\int_0^{0.0001} f_{X|\{d\}} dx \approx 0.8$, such that $d = (\mu_1, \mu_2) = (0.01, 0.002575)$.

   In fact, in monotone and bathtub cases, the beta distribution has a tendency to have higher probability values of events given by even small right neighborhoods of $x = 0$ and small left neighborhoods of $x = 1$. This is the case, that happens frequently for higher values of variances and therefore the choice of a suitable probability distribution is severely handicapped.

   So, if the probabilities of events given by neighborhoods of $x = 0$ and $x = 1$ are relatively high, then the beta distribution can be well simplified and approximated by a simple Bernoulli distribution with the support $\{0, 1\}$.

3. In uni-modal cases of beta distribution for $k_1 > 0$ and $k_2 > 0$, i.e. the cases when the variances are relatively small, the maximum differences are more or less insignificant and enables the usage an appropriate usage of either beta distribution or minimum information distribution for a given set of two moments.

   However, it has to be pointed out that the density function of the beta distribution has zeros (i.e. meets the horizontal axis) at both $x = 0$ and $x = 1$. This is a severe restriction to the generality, as any density function, which should approximate an unknown but existing density function, **should not** meet the **horizontal axis** at its two ends by a strict rule. This loss of generality may not be desirable.

Chapter 10

Needlessness of moments more than two

The **amount of quantitative information** used to determine the probability distribution of a random variable is **defined** by a **finite sequence of moments** of the random variable. This very **definition** of quantitative information has been clearly given and used by several authors, for eg. by the authors of the published papers [3], [4], [59]. However, these authors have not taken the following unforgettably important relevant question under consideration: Do the moments of higher order (owing to the **smallness** of their **ranges of variabilities** (referred to the **definition 5.2.5**) contained by the given quantitative information really **contribute** to the information necessary for the determination of the probability distribution? This very question shall be discussed in this chapter with a set of numerical examples. To my knowledge, not only these authors, there are several other authors, who did seem to have **overlooked** this question and the relevancy of this question. In other words, this question of **contributiveness** shall be discussed in this chapter by illustrating a set of numerical examples.

We shall discuss this **contributiveness**, principally by making use of the **statement of the abstract** of our joint paper (our paper [39]) for to show that, from the **stochastic point of view**, the usage of the **moments of third or higher order are basically useless**. The **smaller** bounded ranges of third and fourth moments, as pictured by the **example 10.0.1**, **speak for themselves**.

We know that the class of uni-modal probability distributions is a subclass

of the class of uni-extremal probability distributions. In fact, uni-modal probability distributions are **very commonly used**. We shall therefore put our assertions by simple numerical examples of uni-modal probability distributions.

10.0.1 A restatement regarding higher moments

In accordance with the minimum information principle, construction of any multi-extremal minimum information probability distribution necessitates the number of moments not less than three. Another word for a **multi-extremal minimum information probability distribution** is a **non-standard minimum information probability distribution**.

From the **stochastic point of view**, the usage of a multi-extremal minimum probability distribution is **rather infrequent** and therefore **rather insignificant**, though it could be a part of a purely mathematical problem.

In this chapter, we shall basically show that we do not need more than two moments to construct a uni-extremal probability distribution, especially from the stochastic point of view. In fact, the utility of the usage of more than two moments is **basically needless** and **useless**.

We know that the class of uni-modal probability distributions is simply a subclass of the class of uni-extremal probability distributions. In fact, since uni-modal probability distributions are **very commonly used** in **Stochastics**, it is therefore **basically enough** for us to show by simple numerical examples that the construction of uni-modal probability distributions **does not necessitate** the usage of more than the **first two moments**. Not only this, the construction of an uni-modal probability distribution by means of **three or more moments** is **contrary to the minimum information principle**.

However, it may be important to state that the usage of a probability distribution from a **stochastic point of view may be** somewhat different from any **mathematical point of view**.

10.0.2 Revisiting the restriction $\frac{\mu_{n-1}^2}{\mu_{n-2}} < \mu_n < \mu_{n-1}$, $n \in \mathbb{N}$

This inequality has already been discussed before in the **chapter 5**. Here, we shall put this inequality to another significant use.

Let us take the reference of our joint paper [39], whose abstract had been given as:

> Any random variable X describing a real phenomenon has necessarily a bounded range of variability implying that the values of the moments determine the probability distribution uniquely. In fact, the range of variability of a random variable restricts the range of the first moment; the value of the first moment limits considerably the range of the second moment; etc. **Thus, any knowledge about the values of lower moments may be used without a further sample for drawing inference on the higher moments**. In this paper we assume without loss of generality that the range of variability of the random variable X is given by the unit interval. Subsequently, the arising restrictions for the three first moments are derived and the implications with respect to uni- modal random variables is investigated and it is shown that **for uni- modal probability distributions the third moment yields only marginal additional information**.

The main substance of this abstract has to be interpreted in form of the following statement:

Statement 10.0.1 (Controlled boundedness of the bounded range of $\mu_n, n \in N$). *We shall restate that the bounded range of μ_n, $n \in N$ is* ***controlled*** *by the moments of lower order, namely $\mu_1, \mu_2, \ldots, \mu_{n-1}$ in the following way:*

- *the bounded range of variability of X, namely the interval $[0, 1]$ restricts the bounded range of the first moment μ_1, the restriction being described by the constraint $0 < \mu_1 < 1$.*
- *the value of the first moment μ_1 restricts the bounded range of the second moment μ_2, the restriction being described by the constraint $\mu_1^2 < \mu_2 < \mu_1$.*
- *in turn, the values of the first μ_1 and the second moment μ_2 restrict the bounded range of the third moment μ_3, the restriction being described by the constraint $\frac{\mu_2^2}{\mu_1} < \mu_3 < \mu_2$.*

- *in turn, the values of the first μ_1, the second μ_2 and the third moment μ_3 restrict the bounded range of the fourth moment μ_4, the restriction being described by the constraint $\frac{\mu_3^2}{\mu_2} < \mu_4 < \mu_3$.*

- *Proceeding exactly in this way, the values of the first $n-1$ moments, namely $\mu_1, \mu_2, \ldots, \mu_{n-1}$ restrict the bounded range of the nth moment μ_n, the restriction being described by the constraint $\frac{\mu_n^2}{\mu_{n-1}} < \mu_{n+1} < \mu_n$.*

For exemplifying our immediately preceding statement **numerically**, we proceed as follows:

Example 10.0.1 (Exemplification of the controlled boundedness of the bounded range of $\mu_n, n \in \mathbb{N}$). *The following simple numerical example is made to start with $\mu_1 = 0.6$*

- *if we take $\mu_1 = 0.6$, then the bounded range of μ_2 would accordingly be $(0.36, 0.6)$.*

- *again, if we take $\mu_1 = 0.6$ and $\mu_2 = 0.48$, then the bounded range of μ_3 would accordingly be $(0.384, 0.48)$.*

- *again, if we take $\mu_1 = 0.6$, $\mu_2 = 0.48$ and $\mu_3 = 0.432$, then the bounded range of μ_4 would accordingly be $(0.3888, 0.432)$.*

- *again, if we take $\mu_1 = 0.6$, $\mu_2 = 0.48$, $\mu_3 = 0.432$ and $\mu_4 = 0.4104$ then the bounded range of μ_5 would accordingly be $(0.38988, 0.4104)$.*

- *again, if we take $\mu_1 = 0.6$, $\mu_2 = 0.48$, $\mu_3 = 0.432$, $\mu_4 = 0.4104$ and $\mu_5 = 0.40014$ then the bounded range of μ_6 would accordingly be $(0.390136, 0.40014)$, etc. etc.*

So, we can well see that the bounded ranges of the moments get **strictly monotonically smaller** and thereby exemplifying our statement 10.0.1 and at the same time exemplifying our previously proven **proposition 5.4.1**. In fact, these bounded ranges get smaller **rather fast**.

10.0.3 Numerical examples

The handled numerical examples in this subsection are **exclusively** referred to our joint paper [39]. In each of these examples, the following are performed:

- An uni-modal probability density is suitably constructed by means of the first **three** moments, only after making absolutely sure that the **third moment does not violate the uni-modal nature** of the probability density. This probability density shall be denoted by $f_{X|\{\mathbf{d}\}}(x)$, $0 \leq x \leq 1$.
- The uni-modal probability density is constructed thereafter by the first two moments of the above case. This probability density shall be denoted by $f_{X|\{d\}}(x)$, $0 \leq x \leq 1$.

In each of the numerical illustrated examples, we have used the following notations:

- $\mu_2^{(U)} = \sigma_{X,U}^2 + \mu_1^2$ as the **limiting** second moment for the **uni- modality** of the probability distribution of X,

  i.e. if μ_2 of the probability distribution **exceeds** $\mu_2^{(U)}$, then the probability distribution is **no longer uni- modal**.
- $\mu_2^{(L)} = \sigma_{X,L}^2 + \mu_1^2$ as the **limiting** second moment for the **bathtub-shapeliness** of the probability distribution of X,

  i.e. if μ_2 of the probability distribution **falls below** $\mu_2^{(L)}$, then the probability distribution is **no longer bathtub- shaped**.
- $\mu_{3L} = \frac{\mu_2^2}{\mu_1}$ as the deduced **greatest lower bound** of μ_3 by the **positive definiteness** of the Hankel matrix (4.25).
- $\mu_{3U} = \mu_2 - \frac{(\mu_1 - \mu_2)^2}{1 - \mu_1}$ as the deduced **least upper bound** of μ_3 by the **positive definiteness** of the Hankel matrix (4.32).
- the third central moment (known to be the **skewness**) of the probability distributions with the probability densities $f_{X|\{\mathbf{d}\}}(x)$, $0 \leq x \leq 1$ and $f_{X|\{d\}}(x)$, $0 \leq x \leq 1$ are denoted by $\widetilde{\zeta}$ and ζ respectively.

Five numerical examples are taken based on different values of μ_1 belonging to the different regions of the open interval $(0, 1)$.

Example 10.0.2 ($\mu_1 \approx 0.5$)**.** *We proceed as follows:*

Usage of three moments*:*

Corresponding to $\mathbf{d} = (\mu_1, \mu_2, \widetilde{\mu}_3) = (0.577305, 0.336507, 0.198052)$*, the probability density is given by*

$$f_{X|\{\mathbf{d}\}}(x) = e^{-100.754+445.2x-613.35x^2+260.5x^3},\ 0 \leq x \leq 1 \tag{10.1}$$

Here,

- $\widetilde{\zeta} = 0.325139$
- *Obviously,* $\widetilde{\mu}_3 = 0.198052$

Usage of two moments*:*

Corresponding to $d = (\mu_1, \mu_2) = (0.577305, 0.336507)$*, the probability density is given by*

$$f_{X|\{d\}}(x) = e^{-49.7041435320758+178.9468012849878x-154.9846279566155x^2},\ 0 \leq x \leq 1 \tag{10.2}$$

Here,

- $\zeta = -1.8767951326058845 \times 10^{-11}$
- $\mu_2^{(U)} = 0.407801 > \mu_2 = 0.336507$ *and therefore the probability density curve is uni- modal.*
- $\mu_2^{(L)} = 0.417901$
- $(\mu_{3L}; \mu_{3U}) = (0.196148; 0.199331)$
- $\mu_3 = 0.197992$

The ***common variance*** *of the two stated probability density curves* $= \mu_2 - \mu_1^2 = \sigma^2 = 0.00322613$*.*

Example 10.0.3 ($\mu_1 \approx 0.7$)**.** *We proceed as follows:*

Usage of three moments*:*

Corresponding to $\mathbf{d} = (\mu_1, \mu_2, \widetilde{\mu}_3) = (0.663605, 0.441038, 0.293562)$*, the probability density is given by*

$$f_{X|\{\mathbf{d}\}}(x) = e^{-772.3106+3005.79x-3774.42x^2+1514.35x^3},\ 0 \leq x \leq 1 \tag{10.3}$$

Here,

- $\widetilde{\zeta} = 0.160539$
- *Obviously,* $\widetilde{\mu}_3 = 0.293562$

Usage of two moments*:*

Corresponding to $d = (\mu_1, \mu_2) = (0.663605, 0.441038)$*, the probability density is given by*

$$f_{X|\{d\}}(x) = e^{-327.6352787696126+995.6919954296214x-750.2143560021559x^2},\ 0 \leq x \leq 1 \tag{10.4}$$

Here,

- $\zeta = -7.612236047794894 \times 10^{-11}$
- $\mu_2^{(U)} = 0.49747 > \mu_2 = 0.441038$ *and therefore the probability density curve is uni- modal.*
- $\mu_2^{(L)} = 0.51714$
- $(\mu_{3L}; \mu_{3U}) = (0.293118; 0.293783)$
- $\mu_3 = 0.29356$

The ***common variance*** *of the two stated probability density curves* $= \mu_2 - \mu_1^2 = \sigma^2 = 0.000666476$*.*

Example 10.0.4 ($\mu_1 \approx 0.3$)**.** *We proceed as follows:*

Usage of three moments*:*

Corresponding to $\mathbf{d} = (\mu_1, \mu_2, \widetilde{\mu}_3) = (0.331225, 0.110203, 0.0368292)$*, the probability density is given by*

$$f_{X|\{\mathbf{d}\}}(x) = e^{-145.002+1005.79x-2024.57x^2+1014.45x^3},\ 0 \leq x \leq 1 \tag{10.5}$$

Here,

- $\widetilde{\zeta} = 0.0669201$
- *Obviously,* $\widetilde{\mu}_3 = 0.0368292$

Usage of two moments*:*

Corresponding to $d = (\mu_1, \mu_2) = (0.331225, 0.110203)$*, the probability density is given by*

$$f_{X|\{d\}}(x) = e^{-108.38720239285139+671.9046009907327x-1014.2721729801988x^2},\ 0 \leq x \leq 1 \tag{10.6}$$

Here,

- $\zeta = -1.1712146443357637 \times 10^{-11}$
- $\mu_2^{(U)} = 0.165568 > \mu_2 = 0.110203$ *and therefore the probability density curve is uni- modal.*
- $\mu_2^{(L)} = 0.1857$
- $(\mu_{3L}; \mu_{3U}) = (0.036666; 0.0371579)$
- $\mu_3 = 0.0368285$

The ***common variance*** *of the two stated probability density curves* $= \mu_2 - \mu_1^2 = \sigma^2 = 0.000492964$.

Example 10.0.5 ($\mu_1 \approx 0.03$)**.** *We proceed as follows:*

Usage of three moments*:*

Corresponding to $\mathbf{d} = (\mu_1, \mu_2, \widetilde{\mu}_3) = (0.0339408, 0.0014847, 0.000075476)$*, the probability density is given by*

$$f_{X|\{\mathbf{d}\}}(x) = e^{1.88237+75.3226x-1259.85x^2+814.794x^3},\ 0 \leq x \leq 1 \tag{10.7}$$

Here,

- $\widetilde{\zeta} = 0.411687$
- *Obviously,* $\widetilde{\mu}_3 = 0.000075476$

Usage of two moments*:*

Corresponding to $d = (\mu_1, \mu_2) = (0.0339408, 0.0014847)$*, the probability density is given by*

$$f_{X|\{d\}}(x) = e^{1.910147349535794+71.87155673715307x-1158.2761849087049x^2},\ 0 \leq x \leq 1 \tag{10.8}$$

Here,

- $\zeta = 0.393576$
- $\mu_2^{(U)} = 0.00180952 > \mu_2 = 0.0014847$ *and therefore the probability density curve is uni- modal.*
- $\mu_2^{(L)} = 0.00241263$
- $(\mu_{3L}; \mu_{3U}) = (0.0000649462; 0.000394289)$
- $\mu_3 = 0.0000753659$

The ***common variance*** *of the two stated probability density curves* $= \mu_2 - \mu_1^2 = \sigma^2 = 0.000332719$*.*

Example 10.0.6 ($\mu_1 \approx 0.84$)**.** *We proceed as follows:*

***Usage of three moments**:*

Corresponding to $\mathbf{d} = (\mu_1, \mu_2, \widetilde{\mu}_3) = (0.840583, 0.712421, 0.608714)$*, the probability density is given by*

$$f_{X|\{\mathbf{d}\}}(x) = e^{-224.507+760.8x-846.15x^2+310.5x^3},\ 0 \leq x \leq 1 \tag{10.9}$$

Here,

- $\widetilde{\zeta} = 0.0985569$
- *Obviously,* $\widetilde{\mu}_3 = 0.608714$

***Usage of two moments**:*

Corresponding to $d = (\mu_1, \mu_2) = (0.840583, 0.712421)$*, the probability density is given by*

$$f_{X|\{d\}}(x) = e^{-51.07566311252618+124.45872541366826x-73.50426482517058x^2},\ 0 \leq x \leq 1 \tag{10.10}$$

Here,

- $\zeta = -0.264148$
- $\mu_2^{(U)} = 0.721086 > \mu_2 = 0.712421$ *and therefore the probability density curve is uni- modal.*
- $\mu_2^{(L)} = 0.739675$
- $(\mu_{3L}; \mu_{3U}) = (0.6038; 0.609387)$
- $\mu_3 = 0.608552$

The ***common variance*** *of the two stated probability density curves* $= \mu_2 - \mu_1^2 = \sigma^2 = 0.0058417$*.*

Moreover, it has to be **importantly** stated that in each of these above examples, the principally important stochastic procedure, namely the prediction procedure (i.e. the procedures for computing **reliable** and **accurate** predictions), which has been modelled by means of both the densities $f_{X|\{\mathbf{d}\}}(x)$, $0 \leq x \leq 1$ and $f_{X|\{d\}}(x)$, $0 \leq x \leq 1$ remains **almost identically the same**. This very thing has been an **important concluding statement** of our paper [39]) too. So, it is **not at all** incorrect to say that the usages these two densities hardly make any noticeable difference.

10.0.4 Conclusions

The following conclusive points follow immediately from the examples:

1. The exemplified probability density curves $f_{X|\{\mathbf{d}\}}(x)$, $0 \leq x \leq 1$ and $f_{X|\{d\}}(x)$, $0 \leq x \leq 1$ almost overlap. So, the differences in their characteristic properties are expectedly barely marginal.

2. The difference between the third moments of both the cases (i.e. the cases of **three** moments and that of **two** moments) is hardly significant. Therefore, the third moment **does not give any significant additional information** addressed to the probability distribution, as far as the uni- modality is concerned.

3. The usage of the third moment for the purpose of constructing a uni- extremal or rather a uni- modal probability distribution would mean nothing, but an **unnecessary additional exponentially higher** amount of the following tasks:

   - Numerical mathematical work and
   - Programming work

   However, if it is a question of constructing a **bi- extremal probability distribution**, then, in accordance with the minimum information principle, this work is unavoidable.

4. One should unforgettably consider the following important points, if one thinks of using the third or higher moments for computing a uni-extremal probability distribution:

- As we have already seen, the bounded range of μ_3 is significantly smaller than the same of μ_2.
- From the **stochastic point of view**, i.e. for the **practical applications**, the true value of μ_3 is, in general, never known. Therefore, an appropriate estimation of μ_3 is necessary and since the **cubic powers** are needed to be computed for this, the computation of an estimated empirical value of μ_3 is significantly more sensitive to estimation errors than the same of μ_2. So, even a **slightest estimation error** in course of estimating μ_3 may cause the violation of the desired uni- extremal property (for eg. the desired unimodal property) of the probability distribution, the probability distribution becomes likely to be bi- extremal. The usage of μ_3 is therefore well beyond practicability.
- The higher moments, as we have seen, have still smaller bounded ranges. The practicability of the usage of higher moments for computation of a uni- extremal probability distribution is therefore well beyond question.

5. The measure of skewness is principally commonly addressed to the unimodal probability distributions. With regard to this, the **third central moment** is **not a good measure of skewness**, as the differences between $\widetilde{\zeta}$ and ζ are **alarmingly high**. The third central moment seems to be **overly sensitive** to the following:
   - The **smallness of the variance**.
   - Even **insignificantly small changes of the third moment**.

The monotonic as well as the uni- extremal (especially the uni- modal) minimum information probability distributions constructed by means of one or two moments are the most essential resources for modelling the **stochastic procedures**.

Part II

Numerical Computations

Chapter 11

The formulation of the ultimate problem

In this dissertation, corresponding to each empirical value of the deterministic variable d_Y, our task shall be to determine the probability distribution of the random variable Y, with subject to the given $\mathcal{X}_Y(\{d_Y\})$. This probability distribution is given exclusively by $\lambda(d_Y)$, which gives $\lambda_1, \lambda_2, \ldots, \lambda_m$ and of course λ_0 as **distributional coefficients** for computer codes.

The derivation of an appropriate probability distribution of Y is based on

- the knowledge about the actual member of the family of probability distributions $\mathbb{P}_m$,
- the minimum information principle, and
- the moments-related representation of the deterministic variable $d_Y = (E[Y], E[Y^2], \ldots, E[Y^m])$, provided $m > 0$. Notably, $m = 0$ means that no information regarding moments is available and the value of λ_0 is solely trivially given without any computational procedure.

The programming work is restricted to the cases of $m \leq 2$. Thus, the formulation of the **ultimate problem of developing computer codes** is the strategy for determining the coefficients $\lambda_i = \lambda_i(d_Y)$ with $i = 1, 2, \ldots, m$ for a empirically given range of variability $\mathcal{X}_Y(\{d_Y\})$ and a given empirical value of the deterministic variable $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)})$ belonging to the parameter space $\mathcal{D}_Y$ for $0 < m \leq 2$. This is the task of this very chapter.

This problem has been addressed in various authors like [25] in the context of MEP-distributions. The computer programs do yield results in most of the cases. However, only in very small cases, the programs may not give the desired solutions. The reason for the arising difficulties will be discussed in due course.

In the subsequent step the applied transformation of the random variable will be discussed. That is, we shall discuss the utility of using the random variable X once again.

11.1 Standardization of the variability

Any random variable $Y|\{d_Y\}$ describing a real-world aspect has necessarily a finite range of variability implying that any continuous approximation should have a bounded range of variability. It follows that any continuous approximation with unbounded range of variability violates the reality and, therefore, is not considered in this dissertation.

Clearly, in view of developing algorithms, it would be beneficial to have only one pre-determined range of variability. Therefore, not the true random variable Y, but a transformed random variable X is used for the specified algorithmic procedures.

We shall denote

$$\begin{array}{rcl} a = a(d_Y) & = & \min\{y \,|\, y \in \mathcal{X}_Y(\{d_Y\})\} \\ b = b(d_Y) & = & \max\{y \,|\, y \in \mathcal{X}_Y(\{d_Y\})\} \end{array} \tag{11.1}$$

Then, the transformed random variable is defined by:

- $X|\{d\} = \frac{Y|\{d_Y\} - a(d_Y)}{b(d_Y) - a(d_Y)}$

  or equivalently

- $Y|\{d_Y\} = a(d_Y) + \Big(b(d_Y) - a(d_Y)\Big)X|\{d\}$.

For the range of variability of the transformed random variable $X|\{d\}$, we immediately obtain

$$\mathcal{X}_X(\{d\}) = \left\{ \begin{array}{lcl} \{0 = x_1, x_2, \ldots, x_n = 1\} & : & \text{for discrete cases} \\ \{x \,|\, 0 \leq x \leq 1\} & : & \text{for continuous cases} \end{array} \right. \tag{11.2}$$

By the transformation (11.1), a random variable is obtained with a range of variability standardized to an unit interval.

This transformation has the following advantages:

1. For the purpose of programming, we need not work with the input variables $a(d_Y)$ and $b(d_Y)$ more than twice: Once during the transformation $X|\{d\} = \frac{Y|\{d_Y\}-a(d_Y)}{b(d_Y)-a(d_Y)}$ and lastly during the reverse transformation $Y|\{d_Y\} = a(d_Y) + (b(d_Y) - a(d_Y))X|\{d\}$ in the final step, after which the results are delivered.

2. Because of the very fact that for μ_i $(i = 1, 2, \ldots, m)$, we already have $1 > \mu_1 > \mu_2 > \ldots > \mu_m > 0$, the computation work for the purpose of computing the $\lambda(d_Y)$ is made simpler.

The transformation (11.1) of the random variable does not only affect the variability function, but also the deterministic variable and the other components of the Bernoulli Space.

11.2 The transformed deterministic variable

The deterministic variable d_Y is represented by the first m moments of Y. The deterministic variable d of the transformed random variable is analogously given by the corresponding moments of X.

$$d = \big(E[X], E[X^2], \ldots, E[X^m]\big) \tag{11.3}$$

or

$$d = (\mu_1, \mu_2, \ldots, \mu_m) \tag{11.4}$$

It is understood, that the influences of d on the random variable X is tantamount to the influence of d_Y on the random variable Y. Therefore, while keeping the following in mind

$$E[(X|\{d\})^i] \text{ and } E[(Y|\{d_Y\})^j],\ \ i, j \in \{1, 2, \ldots, m\} \text{ are connected} \tag{11.5}$$

the relations between the components of d_Y and d for $(0 < m \leq 2)$ are given as

$$E[Y|\{d_Y\}] = a(d_Y) + \Big(b(d_Y) - a(d_Y)\Big)E[X|\{d\}] \tag{11.6}$$

$$E[(Y|\{d_Y\})^2] = a(d_Y)^2 + 2a(d_Y)\Big(b(d_Y) - a(d_Y)\Big)E[X|\{d\}] + \Big(b(d_Y) - a(d_Y)\Big)^2 E\left[(X|\{d\})^2\right] \tag{11.7}$$

and

$$E[X|\{d\}] = \frac{1}{b(d_Y) - a(d_Y)}\Big(E[Y|\{d_Y\}] - a(d_Y)\Big) \tag{11.8}$$

$$E[(X|\{d\})^2] = \frac{1}{\Big(b(d_Y) - a(d_Y)\Big)^2}\Big(E[(Y|\{d_Y\})^2] - 2a(d_Y)E[Y|\{d_Y\}] + a(d_Y)^2\Big) \tag{11.9}$$

From (11.8) and (11.9), we immediately obtain:

$$\mu_1 = \frac{1}{\Big(b(d_Y) - a(d_Y)\Big)}\left[\mu_Y^{(1)} - a(d_Y)\right] \tag{11.10}$$

$$\mu_2 = \frac{1}{\Big(b(d_Y) - a(d_Y)\Big)^2}\left[\mu_Y^{(2)} - 2\mu_Y^{(1)}a(d_Y) + a(d_Y)^2\right] \tag{11.11}$$

Finally, the frequently used real number λ_0 is explained in the following way, the details of which follow immediately in the next section:

- if $m > 0$, then λ_0 is an uniquely determined by an empirical value of d_Y and the range of variability $\mathcal{X}_Y(\{d_Y\})$
- if $m = 0$, then λ_0 is uniquely determined by the range of variability $\mathcal{X}_Y(\{d_Y\})$ only

11.3 The computation strategy

As described so far, the problem is to develop a computer code for the determination of $\lambda(d_Y)$, where d_Y is the moments-related deterministic variable.

To this end $\lambda(d_Y)$ is presented as the solution of a system of simultaneous equations. Basically, there are three cases that are covered.

Moreover, for the sake of simplicity, we shall symbolize $\lambda_i = \lambda_i(d_Y)$, such that $i = 0, 1, \ldots, m$. This basically says that, in our cases we have $\lambda_0 = \lambda_0(d_Y)$, $\lambda_1 = \lambda_1(d_Y)$ and $\lambda_2 = \lambda_2(d_Y)$.

Keeping this in mind, the individual families of probability distributions are described as follows:

- constant family with $f_{Y|\{d_Y\}}(y) = \frac{1}{N}$ for discrete cases, but $\frac{1}{b-a}$ for continuous cases,
- monotone family with $d_Y = (\mu_Y^{(1)})$ and $f_{Y|\{d_Y\}}(y) = e^{\lambda_0+\lambda_1 y}$,
- uni-extremal family with $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$ and $f_{Y|\{d_Y\}}(y) = e^{\lambda_0+\lambda_1 y+\lambda_2 y^2}$.

Now, both for monotone and uni-extremal cases, $\lambda(d_Y)$ can be computed by finding the solutions of (11.17), (11.24), (11.36) and (11.57). In this chapter, we shall discuss about the strategy for solving these equations. Each of these stated equations does have an unique solution. This uniqueness of each of these solutions have been well established.

As a matter of fact, the first theorem of the German mathematician named Felix Hausdorff (stated in [29]) has well supported the existence of the solution of each of the respective systems (11.18), (11.25), (11.37) and (11.58).

Again, it is shown in [19] the range of variability and the values of the moments of a random variable exhibit some strong relations. This have to be taken into account for developing solution algorithms for the above systems of equations.

The formal numerical treatment of solving these equations shall be discussed in the subsequent chapter of Numerical Algorithms in full details.

At the very first step for our strategy for finding the solutions of the aforesaid equations, the following transformation is used:

- $x_j = \frac{y_j-a}{b-a}$ in discrete cases, $j = 1, 2, \ldots, N$, such that $y_1 = a$ and $y_N = b$
- $x = \frac{y-a}{b-a}$ in continuous cases

where, for the sake of simplicity, we have used $a = a(d_Y)$ and $b = b(d_Y)$.

Subsequently,

- in discrete cases, $0 \leq x_j \leq 1 \Leftrightarrow a \leq y_j \leq b$, where the ranges of variability of $X|\{d\}$ and $Y|\{d_Y\}$ are $\mathcal{X}_X(\{d\}) = \{x_1, x_2, \ldots, x_N\}$ and $\mathcal{X}_Y(\{d_Y\}) = \{y_1, y_2, \ldots, y_N\}$ respectively. Here, $x_1 = 0$ and $x_N = 1$

- in continuous cases, $0 \leq x \leq 1 \Leftrightarrow a \leq y \leq b$, where the ranges of variability of $X|\{d\}$ and $Y|\{d_Y\}$ are $\mathcal{X}_X(\{d\}) = [0, 1]$ and $\mathcal{X}_Y(\{d_Y\}) = [a, b]$ respectively.

Therefore, corresponding to a given empirical value of d_Y and the given $\mathcal{X}_Y$, the computation of $\lambda(d_Y)$ can be strategically carried out case by case as follows.

11.4 Discrete uniform probability distribution

11.4.1 The general case with $N > 1$

In this case of the constant family, λ_0 is readily available:

$$\lambda_0 = -\log N \tag{11.12}$$

11.4.2 Special case: $N = 1$

This is simply a trivial case, where $\lambda_0 = 0$. Notably, the variance σ_Y^2 acquires its minimum possible value, i.e. $\sigma_Y^2 = 0$

11.5 Continuous uniform probability distribution

Even in this case of the constant family, λ_0 is readily available:

$$\lambda_0 = -\log(b - a) \tag{11.13}$$

11.6 Discrete monotonic probability distribution

11.6.1 The general case with $N > 2$

For a discrete $Y|\{d_Y\}$, such that $d_Y = (\mu_Y^{(1)})$ or equivalently $d = (\mu_1)$, then the probability mass functions of $Y|\{d_Y\}$ and $X|\{d\}$ are

$$\begin{aligned} f_{Y|\{d_Y\}}(y) &= \frac{e^{\lambda_1 y}}{\sum_{j=1}^{N} e^{\lambda_1 y_j}},\ y \in \mathcal{X}_Y(\{d_Y\}) = \{y_1, y_2, \ldots, y_N\} \\ \text{and } f_{X|\{d\}}(x) &= \frac{e^{\beta x}}{\sum_{j=1}^{N} e^{\beta x_j}},\ x \in \mathcal{X}_X(\{d\}) = \{x_1, x_2, \ldots, x_N\} \end{aligned} \tag{11.14}$$

respectively.

In order to perform this, the following system of two simultaneous equations must to be solved for $\lambda(d_Y) = (\lambda_1)$:

$$\sum_{j=1}^{N} e^{\lambda_0 + \lambda_1 y_j} = 1 \tag{11.15}$$

$$\sum_{j=1}^{N} y_j e^{\lambda_0 + \lambda_1 y_j} = \mu_Y^{(1)} \tag{11.16}$$

which is equivalent to the solution of the following equation in λ_1:

$$\mu_Y^{(1)} = \frac{\sum_{j=1}^{N} y_j e^{\lambda_1 y_j}}{\sum_{j=1}^{N} e^{\lambda_1 y_j}} \tag{11.17}$$

Therefore, in order to compute the desired $\lambda(d_Y) = (\lambda_1)$, the solution of (11.17) is yielded by the solving the following equation in β at first:

$$\mu_1 = \frac{\sum_{j=1}^{N} x_j e^{\beta x_j}}{\sum_{j=1}^{N} e^{\beta x_j}} \tag{11.18}$$

and only thereafter we arrive at

- $\lambda_1 = \frac{\beta}{b-a}$
- $\lambda_0 = -\log\left(\sum_{j=1}^{N} e^{\lambda_1 y_j}\right)$

11.6.2 The case for constancy

In case $\mu_Y^{(1)} = \frac{\sum_{j=1}^{N} y_j}{N}$ (or equivalently $\mu_1 = \frac{\sum_{j=1}^{N} x_j}{N}$), then the probability distribution is an uniform distribution, such that

- $\lambda_0 = -\log N$
- $\lambda_1 = 0$

11.6.3 The trivial case: $N = 1$

In this case, the input of $\mu_Y^{(1)}$ is either redundant or inconsistent. The program therefore gives the result that is independent of the given $\mu_Y^{(1)}$, so that

- $\lambda_0 = 0$
- $\lambda_1 = 0$

11.6.4 Special case: $N = 2$

Here, $\mathcal{X}_Y(\{d_Y\}) = \{a, b\} \Leftrightarrow \mathcal{X}_X(\{d\}) = \{0, 1\}$ and therefore with subject to $0 < \mu_1 < 1$, we get

$$\begin{aligned} P_{X|\{d\}}(\{0\}) &= 1 - \mu_1 \\ P_{X|\{d\}}(\{1\}) &= \mu_1 \end{aligned}$$

which leads us to

$$P_{X|\{d\}}(\{x\}) = e^{\alpha+\beta x} \text{ for } x \in \mathcal{X}_X(\{d\}) \tag{11.19}$$

such that

- $\alpha = \log(1-\mu_1)$ and
- $\beta = -\log\left(\frac{1}{\mu_1} - 1\right)$

Now, by comparing the coefficients of y in the following to get λ_0 and λ_1,

$$e^{\lambda_0+\lambda_1 y} = e^{\alpha+\beta x} = e^{\alpha+\beta\left(\frac{y-a}{b-a}\right)}$$

we easily get

$$P_{Y|\{d_Y\}}(\{y\}) = e^{\lambda_0+\lambda_1 y} \text{ for } y \in \mathcal{X}_Y(\{d_Y\}) \tag{11.20}$$

such that

- $\lambda_0 = \log\left(\frac{b-\mu_Y^{(1)}}{b-a}\right) + \frac{a}{b-a}\log\left(\frac{b-\mu_Y^{(1)}}{\mu_Y^{(1)}-a}\right)$ and
- $\lambda_1 = -\frac{1}{b-a}\log\left(\frac{b-\mu_Y^{(1)}}{\mu_Y^{(1)}-a}\right)$

Notably, the variance σ_Y^2 acquires its maximum possible value, i.e. $\sigma_Y^2 = (\mu_Y^{(1)} - a)(b - \mu_Y^{(1)})$

11.7 Continuous monotonic probability distribution

11.7.1 The general case

For a continuous $Y|\{d_Y\}$, such that $d_Y = (\mu_Y^{(1)})$ or equivalently $d = (\mu_1)$, then the probability density functions of $Y|\{d_Y\}$ and $X|\{d\}$ are

$$\begin{aligned} f_{Y|\{d_Y\}}(y) &= \frac{e^{\lambda_1 y}}{\int_a^b e^{\lambda_1 y} dy},\ y \in \mathcal{X}_Y(\{d_Y\}) = [a,b] \\ \text{and } f_{X|\{d\}}(x) &= \frac{e^{\beta x}}{\int_0^1 e^{\beta x} dx},\ x \in \mathcal{X}_X(\{d\}) = [0,1] \end{aligned} \tag{11.21}$$

respectively.

In order to perform this,, the following system of two simultaneous equations must to be solved for $\lambda(d_Y) = (\lambda_1)$:

$$\int_a^b e^{\lambda_0+\lambda_1 y} dy = 1 \tag{11.22}$$

$$\int_a^b y e^{\lambda_0+\lambda_1 y} dy = \mu_Y^{(1)} \tag{11.23}$$

which is equivalent to the solution of the following equation in λ_1:

$$\mu_Y^{(1)} = \frac{\int_a^b y e^{\lambda_1 y} dy}{\int_a^b e^{\lambda_1 y} dy} \tag{11.24}$$

Therefore, in order to compute the desired $\lambda(d_Y) = (\lambda_1)$, the solution of (11.24) is yielded by the solving of the following equation in β at first:

$$\mu_1 = \frac{\int_0^1 x e^{\beta x} dx}{\int_0^1 e^{\beta x} dx} \tag{11.25}$$

and only thereafter we arrive at

- $\lambda_1 = \frac{\beta}{b-a}$
- $\lambda_0 = -\log\left(\int_a^b e^{\lambda_1 y} dy\right)$

11.7.2 The case for constancy

This is the case, when $\mu_Y^{(1)} = \frac{a+b}{2}$ or equivalently $\mu_1 = \frac{1}{2}$.

Here,

- $\lambda_1 = 0$
- $\lambda_0 = -\log(b-a)$

11.8 Usage of the standard normal density

Soon we shall come across through certain cases, when the probability density function

$$f_{X|\{\widetilde{d}\}}(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2}}{\int\limits_0^1 e^{-\frac{1}{2}\left(\frac{t-\mu_1}{\sigma}\right)^2} dt}, \quad 0 \leq x \leq 1,\ 0 < \mu_1 < 1,\ 0 < \sigma < \sqrt{\mu_1(1-\mu_1)} \tag{11.26}$$

determined by $\widetilde{d} = (\widetilde{\mu}_1, \widetilde{\mu}_2)$ (such that $\widetilde{\sigma}^2 = \widetilde{\mu}_2 - \widetilde{\mu}_1^2$) of the continuous random variable X can be used as an **utilizable approximation** of the probability density function $f_{X|\{d\}}(x) = \frac{e^{\beta x+\gamma x^2}}{\int\limits_0^1 e^{\beta t+\gamma t^2} dt}, 0 \leq x \leq 1$ determined by $d = (\mu_1, \mu_2)$ where μ_1 and $\sigma = \sqrt{\mu_2 - \mu_1^2}$ are the **user-given** mean and standard deviation respectively. $\widetilde{\mu}_1$ and $\widetilde{\sigma}$ are denoted as the actual mean and the standard deviation of the probability density (11.26). Quite obviously, $\mu_1 \neq \widetilde{\mu}_1$ and $\sigma \neq \widetilde{\sigma}$ are the cases, in general.

This **approximative approach** investigates the cases, how the conditions $\mu_1 \approx \widetilde{\mu}_1$ and $\sigma \approx \widetilde{\sigma}$ help to to make sure that the probability density of X given by (11.26) approximates the **minimum information** probability density $f_{X|\{d\}}(x)$ of X with respect to the predetermined mean μ_1 and variance σ^2 well enough.

In this section, with the help of the probability density function of the standard normal distribution (i.e. (11.26)), we shall find out that condition, under which

- the usage of the approximated probability distribution (11.26) of X is justified.

- the expressions $|\mu_1 - \widetilde{\mu}_1|$ and $|\sigma - \widetilde{\sigma}|$, which are formally derived in one of the subsequent subsections, **decrease rapidly** with the **decrease** in σ. Therefore, the **goodness** of the aforesaid approximation **increases** with the **decrease** in σ.

According to the knowledge of the standard normal probability density func-

tion, we have

$$\frac{1}{\sqrt{2\pi}}\int_{-\ell}^{\ell} e^{-\frac{x^2}{2}}dx = 1-\epsilon(\ell) \tag{11.27}$$

where $\epsilon(\ell)$ is a function of ℓ, which tends to zero rapidly with the increase in ℓ. As a matter of fact, for $\ell = 3$, $\epsilon(\ell) = 0.0027$ & for $\ell = 4$, $\epsilon(\ell) = 0.00006$, etc.

By rewriting the integral $\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2}dx$ (which is a part of (11.26)) and by the knowledge of the standard normal probability distribution, with regard to $0 < \mu_1 < 1$, we get

$$\frac{1}{\sigma\sqrt{2\pi}}\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\frac{\mu_1}{\sigma}}^{\frac{1-\mu_1}{\sigma}} e^{-\frac{x^2}{2}}dx \overset{\sigma\to 0}{\longrightarrow} 1 \tag{11.28}$$

Therefore, it is evidently clear, that our approximation improves with the decrease in σ.

Keeping this in mind, by comparing the right hand side integral of the (11.28) with the left hand side integral of the (11.27), we can use $\frac{1}{\sqrt{2\pi}}\int_{-\ell}^{\ell} e^{-\frac{x^2}{2}}dx$ as a reliable **lower bound** of $\frac{1}{\sqrt{2\pi}}\int_{-\frac{\mu_1}{\sigma}}^{\frac{1-\mu_1}{\sigma}} e^{-\frac{x^2}{2}}dx$. This leads us to lead the following:

- $\ell < \frac{1-\mu_1}{\sigma}$
- $-\ell > -\frac{\mu_1}{\sigma}$

which means

- $\sigma < \frac{1-\mu_1}{\ell}$
- $\sigma < \frac{\mu_1}{\ell}$

and therefore, by taking $\mu_D = \min(\mu_1, 1-\mu_1)$, we get the required **approximating condition** as

$$\sigma < \frac{\mu_D}{\ell} \tag{11.29}$$

Furthermore, cases may arise, when X must be a discrete random variable instead of being continuous. Even in such cases, we can make use of the standard normal distribution effectively. This is done in the following manner:

Let the bounded interval $[0,1]$ be divided into N parts by means of the finite set of points $x_1, x_2, \ldots, x_N$, such that $0 = x_1 < x_2 < \ldots < x_N = 1$. Then by taking $\delta_r = x_{r+1} - x_r$, where $r = 1, 2, \ldots, N-1$, for a large N, we get the expression of the probability that X lies in between c and d $(0 \leq c < d \leq 1)$ as

$$\begin{aligned}
\int_c^d f_{X|\{\widetilde{d}\}}(x)dx &= \frac{\int_c^d e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx}{\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx} \\
&= \frac{\int_{x_j}^{x_k} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx}{\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx} = \frac{\lim\limits_{\delta\to 0}\sum\limits_{i=j}^{k} \delta_i e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}}{\lim\limits_{\delta\to 0}\sum\limits_{i=1}^{N} \delta_i e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}} \\
&= \frac{\lim\limits_{(k-j)\to\infty}\sum\limits_{i=j}^{k} \delta_i e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}}{\lim\limits_{N\to\infty}\sum\limits_{i=1}^{N} \delta_i e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}} \\
&\approx \frac{\sum\limits_{i=j}^{k} \delta e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}}{\sum\limits_{i=1}^{N} \delta e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}} \\
&= \frac{\sum\limits_{i=j}^{k} e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}}{\sum\limits_{i=1}^{N} e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}} \qquad (11.30)
\end{aligned}$$

such that

- we have taken $c = x_j$ and $d = x_k$ where $1 \leq j < k \leq N$
- δ is called the norm of subdivision, defined by $\delta = \max\limits_{r\in\{1,2,\ldots,N-1\}} \delta_r$. Therefore $N \to \infty \Leftrightarrow \delta \to 0$

It is absolutely clear, that in the **penultimate** step, the proposed approximation is only admissible, if N is **sufficiently large** and δ is **sufficiently small**.

In particular, if δ_r is constant for all values of r, then subsequently we have $N\delta$ =1 for every choice of N. This is just the case, when $x_1, x_2, \ldots, x_N$ are in arithmetic progression.

This leads us to assert our probability mass function of X, with subject to the user-given mean μ_1 and standard deviation σ, as

$$f_{X|\{\widetilde{d}\}}(x_j) = \frac{e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}{\sum\limits_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}, \quad j = 1, 2, \ldots, N \tag{11.31}$$

such that $0 = x_1 < \ldots < x_N = 1,\ 0 < \mu_1 < 1,\ 0 < \sigma < \sqrt{\mu_1(1-\mu_1)}$.

Obviously, for large values of N, the same approximating condition (11.29) can be harmlessly used, in cases, when $x_1, x_2, \ldots, x_N$ are in arithmetic progression.

Moreover, my programming experience clearly shows, that the quality approximating discrete probability distribution becomes good, if $N \geq 100$ and δ is relatively small.

In the discrete case, the derived expressions $|\widetilde{\mu}_1 - \mu_1|$ and $|\widetilde{\sigma} - \sigma|$ given in one of the subsequent subsections show, that they decrease speedily with the decrease in σ in many cases. However, the decreasing nature of $|\widetilde{\sigma} - \sigma|$ with the monotonic decrease in σ is additionally **heavily** dependent on the predeterminately given finite support $\{0 = x_1, x_2, \ldots, x_N = 1\}$ as an input.

When X is continuous, the fulfillment of the **approximating condition (11.29)** confirms the the smallness simultaneously of **both** the expressions of $|\widetilde{\mu}_1 - \mu_1|$ and $|\widetilde{\sigma} - \sigma|$. But, if X is discrete, the condition (11.29) may not be enough to confirm smallness simultaneously of both $|\widetilde{\mu}_1 - \mu_1|$ and $|\widetilde{\sigma} - \sigma|$.

The simultaneous smallness of both the expressions $|\widetilde{\mu}_1 - \mu_1|$ and $|\widetilde{\sigma} - \sigma|$ is **decisive** for the justification of **usage** of the **approximating** probability mass function (11.31) in case of the **discrete** X and of **usage** of the approximating probability density (11.26) in case of the **continuous** X.

So, by keeping the smallness of **both** $|\widetilde{\mu}_1 - \mu_1|$ and $|\widetilde{\sigma} - \sigma|$ in mind (i.e. by keeping $\mu_1 \approx \mu_x$ and $\sigma \approx \sigma_x$ in mind), **the user has to decide**, whether to opt for **inputting $\widetilde{\mu}_1$ and $\widetilde{\sigma}$ as input values instead of μ_1 and σ respectively or not. Of course, the programmer has the responsibility to structure his programs in terms of the feasibilities in this regard (i.e. with regard to this aforesaid smallness) and this shall be discussed in the coming subsections (both in discrete and continuous cases), both of which are named as *Uni-extremal probability distribution with a small variance*** belonging to the coming chapter of Numerical algorithms.

11.9 Discrete uni-extremal probability distribution

11.9.1 The general case with $N > 3$

For a discrete $Y|\{d_Y\}$, such that $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$ or equivalently $d = (\mu_1, \mu_2)$, then the probability mass functions of $Y|\{d_Y\}$ and $X|\{d\}$ are

$$\begin{aligned} f_{Y|\{d_Y\}}(y) &= \frac{e^{\lambda_1 y + \lambda_2 y^2}}{\sum_{j=1}^{N} e^{\lambda_1 y_j + \lambda_2 y_j^2}},\ y \in \mathcal{X}_Y(\{d_Y\}) = \{y_1, y_2, \ldots, y_N\} \\ \text{and } f_{X|\{d\}}(x) &= \frac{e^{\beta x + \gamma x^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}},\ x \in \mathcal{X}_X(\{d\}) = \{x_1, x_2, \ldots, x_N\} \end{aligned} \tag{11.32}$$

respectively.

In order to perform this, the following system of three simultaneous equations must to be solved for $\lambda(d_Y) = (\lambda_1, \lambda_2)$:

$$\sum_{j=1}^{N} e^{\lambda_0 + \lambda_1 y_j + \lambda_2 y_j^2} = 1 \tag{11.33}$$

$$\sum_{j=1}^{N} y_j e^{\lambda_0 + \lambda_1 y_j + \lambda_2 y_j^2} = \mu_Y^{(1)} \tag{11.34}$$

$$\sum_{j=1}^{N} y_j^2 e^{\lambda_0 + \lambda_1 y_j + \lambda_2 y_j^2} = \mu_Y^{(2)} \tag{11.35}$$

which is equivalent to the solution of the following system of simultaneous equations in λ_1 and λ_2:

$$\begin{cases} \mu_Y^{(1)} &= \dfrac{\sum_{j=1}^{N} y_j e^{\lambda_1 y_j + \lambda_2 y_j^2}}{\sum_{j=1}^{N} e^{\lambda_1 y_j + \lambda_2 y_j^2}} \\ \mu_Y^{(2)} &= \dfrac{\sum_{j=1}^{N} y_j^2 e^{\lambda_1 y_j + \lambda_2 y_j^2}}{\sum_{j=1}^{N} e^{\lambda_1 y_j + \lambda_2 y_j^2}} \end{cases} \tag{11.36}$$

Therefore, in order to compute the desired $\lambda(d_Y) = (\lambda_1, \lambda_2)$, the solution of (11.36) is yielded by the solving of the following simultaneous system of equations in β and γ at first:

$$\begin{cases} \mu_1 = \frac{\sum_{j=1}^{N} x_j e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} \\ \mu_2 = \frac{\sum_{j=1}^{N} x_j^2 e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_i j^2}} \end{cases} \tag{11.37}$$

and only thereafter we arrive at

- $\lambda_1 = \frac{\beta(b-a)-2a\gamma}{(b-a)^2}$
- $\lambda_2 = \frac{\gamma}{(b-a)^2}$
- $\lambda_0 = -\log\left(\sum_{j=1}^{N} e^{\lambda_1 y_j + \lambda_2 y_j^2}\right)$

  For the purpose of handling the overflow computing errors, λ_0 is computed according to the following rule:

  - if $(\beta + \gamma < 709)$ then

    $\lambda_0 = -\frac{\beta a}{b-a} + \frac{\gamma a^2}{(b-a)^2} - \log\left(\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}\right)$
  - if $(\beta + \gamma \geq 709)$ then

    $\lambda_0 = -\beta - \gamma - \frac{\beta a}{b-a} + \frac{\gamma a^2}{(b-a)^2} - \log\left(\sum_{j=1}^{N} e^{\beta(x_j-1)+\gamma(x_j^2-1)}\right)$

11.9.2 The case for constancy

In case $\mu_Y^{(1)} = \frac{\sum_{j=1}^{N} y_j}{N}$ and $\mu_Y^{(2)} = \frac{\sum_{j=1}^{N} y_j^2}{N}$ occur simultaneously (or equivalently $\mu_1 = \frac{\sum_{j=1}^{N} x_j}{N}$ and $\mu_2 = \frac{\sum_{j=1}^{N} x_j^2}{N}$ occur simultaneously), then the probability distribution is an uniform distribution, such that

- $\lambda_0 = -\log N$
- $\lambda_1 = 0$
- $\lambda_2 = 0$

11.9.3 The monotonic case

Let β_0 be the (unique) solution of the following equation in β:

$$\mu_1 = \frac{\sum_{j=1}^{N} x_j e^{\beta x_j}}{\sum_{j=1}^{N} e^{\beta x_j}}$$

Then, if so happens that β_0 fulfills the following equality:

$$\mu_2 = \frac{\sum_{j=1}^{N} x_j^2 e^{\beta_0 x_j}}{\sum_{j=1}^{N} e^{\beta_0 x_j}}$$

then the probability distribution is a monotonic probability distribution, which means

- $\lambda_0 = -\log\left(\sum_{j=1}^{N} e^{\frac{\beta_0}{b-a} y_j}\right)$
- $\lambda_1 = \frac{\beta_0}{b-a}$
- $\lambda_2 = 0$

Notably, if $\beta_0 = 0$ then the probability distribution is uniform.

11.9.4 The trivial cases: $N = 1$ and $N = 2$

Subcase 1: $N = 1$. In this case, each of the inputs of $\mu_Y^{(1)}$ and $\mu_Y^{(2)}$ (or equivalently σ_Y^2 can be given instead of $\mu_Y^{(2)}$) is either redundant or inconsistent. The program therefore gives the result that is independent of both the given $\mu_Y^{(1)}$ and $\mu_Y^{(2)}$, so that

- $\lambda_0 = 0$
- $\lambda_1 = 0$
- $\lambda_2 = 0$

Subcase 2: $N = 2$. In this case, the input of $\mu_Y^{(1)}$ is necessary, but the input of $\mu_Y^{(2)}$ (or equivalently σ_Y^2) is either redundant or inconsistent. The program therefore gives the result that is consistent with the predetermined $\mu_Y^{(1)}$, but is independent of the given $\mu_Y^{(2)}$, so that

- $\lambda_0 = \log\left(\frac{b-\mu_Y^{(1)}}{b-a}\right) + \frac{a}{b-a}\log\left(\frac{b-\mu_Y^{(1)}}{\mu_Y^{(1)}-a}\right)$ and
- $\lambda_1 = -\frac{1}{b-a}\log\left(\frac{b-\mu_Y^{(1)}}{\mu_Y^{(1)}-a}\right)$
- $\lambda_2 = 0$

11.9.5 Uni-extremal probability distribution with a small variance

Let $\mu_Y^{(1)}$ and σ_Y^2 be the respective **user-given** mean and variance of the uni-extremal probability distribution of Y with the given finite support $\{y_1, y_2, \ldots, y_N\}$.

Let the random variable Y be assumed to follow an approximated probability distribution with the support $\{y_1, y_2, \ldots, y_N\}$, whose probability mass function $f_{Y|\{\widetilde{d}_Y\}}(y_j)$, $j = 1, 2, \ldots, N$ determined by $\widetilde{d}_Y = (\widetilde{\mu}_Y^{(1)}, \widetilde{\mu}_Y^{(2)})$, such

that $\widetilde{\sigma}_Y^2 = \widetilde{\mu}_Y^{(2)} - \left(\widetilde{\mu}_Y^{(1)}\right)^2$, is given by

$$\begin{aligned} f_{Y|\{\widetilde{d}_Y\}}(y_j) &= \frac{e^{-\frac{1}{2}\left(\frac{y_j-\mu_Y^{(1)}}{\sigma_Y}\right)^2}}{\sum\limits_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{y_j-\mu_Y^{(1)}}{\sigma_Y}\right)^2}}, \quad a = y_1 < y_2 < \ldots < y_N = b \\ &= \frac{K}{\sigma_Y\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j-\mu_Y^{(1)}}{\sigma_Y}\right)^2}, \quad j = 1, 2, \ldots, N \end{aligned} \tag{11.38}$$

By the linear transformation $X = \frac{Y-a}{b-a}$, the probability mass function $f_{X|\{\widetilde{d}\}}(x_j)$, $j = 1, 2, \ldots, N$ of the random variable X is given by

$$f_{X|\{\widetilde{d}\}}(x_j) = \frac{K}{\sigma_Y\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}, \quad 0 = x_1 < x_2 < \ldots < x_N = 1 \tag{11.39}$$

$\left(\because \frac{y_j-\mu_Y^{(1)}}{\sigma_Y} = \frac{x_j-\mu_1}{\sigma}, \text{ for } j \in \{1, 2, \ldots, N\}\right)$, such that for $x_j = \frac{y_j-a}{b-a}$, together with $f_{X|\{\widetilde{d}\}}(x_j) = f_{Y|\{\widetilde{d}_Y\}}(y_j)$, $j = 1, 2, \ldots, N$, we have

- $\mu_1 = \frac{\mu_Y^{(1)}-a}{b-a}$
- $\sigma = \frac{\sigma_Y}{b-a}$
- $K = \frac{\sigma_Y\sqrt{2\pi}}{\sum\limits_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{y_j-\mu_Y^{(1)}}{\sigma_Y}\right)^2}} = \frac{\sigma_Y\sqrt{2\pi}}{\sum\limits_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}} = \frac{(b-a)\sigma\sqrt{2\pi}}{\sum\limits_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}$

We shall make use of the values of μ_1 and σ to judge, whether the uni-extremal probability distribution (11.38) of Y can be taken for the **approximating** probability distribution of Y with small variance (having mean $\widetilde{\mu}_Y$ and standard deviation $\widetilde{\sigma}_Y$, such that $\widetilde{\mu}_Y \approx \mu_Y$ and $\widetilde{\sigma}_Y \approx \mu_Y$).

Therefore, let us **derive** and **examine** the **smallness** (in terms of magnitude) of the values of $\epsilon_{\mu_1} = \widetilde{\mu}_1 - \mu_1$ and $\epsilon_\sigma = \widetilde{\sigma} - \sigma$, where $\widetilde{\mu}_1$ and $\widetilde{\sigma}$ are the mean and standard deviation of the probability mass function of X given by (11.39) respectively. Here,

$$\begin{aligned}\widetilde{\mu}_1 &= \frac{\sum_{j=1}^{N} x_j e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}{\sum_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}} = \frac{\sum_{j=1}^{N} (x_j-\mu_1) e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}{\sum_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}} + \mu_1 \\ &= \frac{K}{\sigma_Y\sqrt{2\pi}} \sum_{j=1}^{N} (x_j-\mu_1) e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2} + \mu_1\end{aligned} \tag{11.40}$$

giving

$$\epsilon_{\mu_1} = \widetilde{\mu}_1 - \mu_1 = \frac{K}{\sigma_Y\sqrt{2\pi}} \sum_{j=1}^{N} (x_j-\mu_1) e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2} \tag{11.41}$$

$$\begin{aligned}\widetilde{\sigma}^2 &= \frac{\sum_{j=1}^{N} (x_j-\widetilde{\mu}_1)^2 e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}{\sum_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}} = \frac{\sum_{j=1}^{N} (x_j-\mu_1+\mu_1-\widetilde{\mu}_1)^2 e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}{\sum_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}} \\ &= \frac{\sum_{j=1}^{N} (x_j-\mu_1-\epsilon_{\mu_1})^2 e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}{\sum_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}} \\ &= \frac{\sum_{j=1}^{N} \left\{(x_j-\mu_1)^2 - 2\epsilon_{\mu_1}(x_j-\mu_1) + \epsilon_{\mu_1}^2\right\} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}{\sum_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}} \\ &= \frac{K}{\sigma_Y\sqrt{2\pi}} \sum_{j=1}^{N} (x_j-\mu_1)^2 e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2} - \epsilon_{\mu_1}^2\end{aligned} \tag{11.42}$$

giving

$$\epsilon_\sigma = \sqrt{\frac{K}{\sigma_Y\sqrt{2\pi}} \sum_{j=1}^{N} (x_j-\mu_1)^2 e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2} - \epsilon_{\mu_1}^2} - \sigma \tag{11.43}$$

Hence, the expressions (11.41) and (11.43) of ϵ_{μ_1} and ϵ_σ respectively can also be written as

$$\epsilon_{\mu_1} = \frac{\sum_{j=1}^{N}(x_j-\mu_1)e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}{\sum_{j=1}^{N}e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}} \tag{11.44}$$

$$\epsilon_\sigma = \sqrt{\frac{\sum_{j=1}^{N}(x_j-\mu_1)^2e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}}{\sum_{j=1}^{N}e^{-\frac{1}{2}\left(\frac{x_j-\mu_1}{\sigma}\right)^2}} - \epsilon_{\mu_1}^2 - \sigma} \tag{11.45}$$

and thus, ϵ_{μ_1} and ϵ_σ are expected to decrease rapidly with the decrease in σ.

As we have already discussed before, apart from the fact that N must be reasonably large and $\delta = \max_{j\in\{1,2,\ldots,N-1\}}(x_{j+1}-x_j)$ must be reasonably small, the **additional** probable approximating condition, for which (11.39) is acceptable as the probability distribution of X with subject to the user-given mean μ_1 and the standard deviation σ reads $\sigma < \frac{\mu_D}{\ell}$, where $\mu_D = \min(\mu_1, 1-\mu_1)$.

Our above usage of the word **probable** means, the aforesaid approximating condition may be of a big help, but not the sufficient condition to ensure the simultaneous smallness of ϵ_{μ_1} and ϵ_σ.

Consequently, this probable approximating condition leads to the acceptance of (11.38) to be the probability distribution of Y, with subject to the user-given mean μ_Y and the standard deviation σ_Y.

Therefore, we can come to the following conclusion:

(11.38) is accepted to be the approximated probability distribution of Y, if the following cases are under consideration:

- $\sigma < \frac{\mu_D}{\ell}$. In our cases, ℓ is made to range from 3.5 to 4.5 for a sufficiently large N. This helps a lot, but not the sufficient condition
- for a preassigned positive number ϵ, if we have $|\epsilon_{\mu_1}| < \epsilon$ and $|\epsilon_\sigma| < \epsilon$. In our cases, ϵ is made to range from 10^{-10} to 10^{-9}. This is the usable condition

Now, let us come to the determination of λ values. They read as follows:

- $\lambda_0 = \log\left(\frac{K}{\sigma_Y\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{\mu_Y}{\sigma_Y}\right)^2$
- $\lambda_1 = \frac{\mu_Y}{\sigma_Y^2}$
- $\lambda_2 = -\frac{1}{2\sigma_Y^2}$

11.9.6 Special case: $N = 3$

Here, $\mathcal{X}_Y(\{d_Y\}) = \{a, \widehat{y}, b\} \Leftrightarrow \mathcal{X}_X(\{d\}) = \{0, \widehat{x}, 1\}$ and therefore with subject to $0 < \mu_1 < 1$ and $\mu_1^2 < \mu_2 < \mu_1$ (i.e. equivalently $0 < \sigma^2 < \mu_1(1-\mu_1)$), we set

$$\begin{aligned} P_{X|\{d\}}(\{0\}) &= 1 - p - q \\ P_{X|\{d\}}(\{\widehat{x}\}) &= p \\ P_{X|\{d\}}(\{1\}) &= q \end{aligned}$$

Therefore, we need to solve the following system of simultaneous equations in p and q, namely

$$\begin{aligned} \mu_1 &= x_1(1-p-q) + x_2 p + x_3 q \\ \mu_2 &= x_1^2(1-p-q) + x_2^2 p + x_3^2 q \end{aligned}$$

Because of $x_1 = 0$, $x_2 = \widehat{x}$ and $x_3 = 1$, the above system of simultaneous equations gets simplified to

$$\begin{aligned} \mu_1 &= \widehat{x}p + q \\ \mu_2 &= \widehat{x}^2 p + q \end{aligned}$$

and on solving it, we get the following

- $$1 - p - q = 1 - \mu_1 - \frac{\mu_1 - \mu_2}{\widehat{x}} \tag{11.46}$$
- $$p = \frac{\mu_1 - \mu_2}{\widehat{x}(1-\widehat{x})} \tag{11.47}$$

- $$q = \frac{\mu_2 - \mu_1 \widehat{x}}{1 - \widehat{x}} \tag{11.48}$$

which leads us to

$$P_{X|\{d\}}(\{x\}) = e^{\alpha+\beta x+\gamma x^2} \text{ for } x \in \mathcal{X}_X(\{d\}) \tag{11.49}$$

such that

- $e^{\alpha} = 1 - p - q$
- $e^{\alpha+\beta\widehat{x}+\gamma\widehat{x}^2} = p$ and
- $e^{\alpha+\beta+\gamma} = q$

thereby giving

- $\alpha = \log(1 - p - q)$
- $\beta = \frac{\log\left(\frac{p}{1-p-q}\right) - \widehat{x}^2 \log\left(\frac{q}{1-p-q}\right)}{\widehat{x}(1-\widehat{x})}$ and
- $\gamma = \frac{\widehat{x} \log\left(\frac{q}{1-p-q}\right) - \log\left(\frac{p}{1-p-q}\right)}{\widehat{x}(1-\widehat{x})}$

Now, by comparing the coefficients of y in the following to get λ_0, λ_1 and λ_2,

$$e^{\lambda_0+\lambda_1 y+\lambda_2 y^2} = e^{\alpha+\beta x+\gamma x^2} = e^{\alpha+\beta\left(\frac{y-a}{b-a}\right)+\gamma\left(\frac{y-a}{b-a}\right)^2}$$

we easily get

$$P_{Y|\{d_Y\}}(\{y\}) = e^{\lambda_0+\lambda_1 y+\lambda_2 y^2} \text{ for } y \in \mathcal{X}_Y(\{d_Y\}) \tag{11.50}$$

such that

- $\lambda_0 = \alpha - \frac{\beta a}{b-a} + \frac{\gamma a^2}{(b-a)^2}$
- $\lambda_1 = \frac{\beta}{b-a} - \frac{2a\gamma}{(b-a)^2}$ and
- $\lambda_2 = \frac{\gamma}{(b-a)^2}$

Now, as we have already discussed about the **existence of certain discrete probability distributions** in certain cases with subject to the predetermined $\mathcal{X}_Y(\{d_Y\})$ (or equivalently $\mathcal{X}_X(\{d\})$). This is exactly one of such cases in which the existence is in question. This is a typical analytical example of this, which we shall discuss here.

The existence **imperatively** necessitates the simultaneous fulfillment of the following:

$$p > 0, \;\; q > 0, \text{ and } 1 - p - q > 0 \tag{11.51}$$

Here, we see that

1. By (11.47), $p > 0 \Leftrightarrow \mu_1 > \mu_2$, which always holds true.
2. By (11.48), $q > 0 \Leftrightarrow \mu_2 > \mu_1\widehat{x} \Leftrightarrow \mu_2 > \mu_1^2 + \mu_1(\widehat{x} - \mu_1)$
3. By (11.46), $1-p-q > 0 \Leftrightarrow 1-\mu_1 > \frac{\mu_1-\mu_2}{\widehat{x}} \Leftrightarrow \mu_2 > \mu_1^2+(1-\mu_1)(\mu_1-\widehat{x})$

and therefore, by summarizing all the necessities of the aforesaid existence, we arrive at the following **condition for the existence imposed on $\mathcal{X}_Y(\{d_Y\})$ (or equivalently on $\mathcal{X}_X(\{d\})$)**:

$$\mu_2 > \max\left\{\mu_1^2 + \mu_1(\widehat{x} - \mu_1) \;\; , \;\; \mu_1^2 + (1 - \mu_1)(\mu_1 - \widehat{x})\right\} \tag{11.52}$$

Hence, we arrive at the following conclusions:

1. For $\mu_1 > \widehat{x}$,
   - The probability distribution does not exist, if $\mu_1^2 < \mu_2 \leq \mu_1^2 + (1 - \mu_1)(\mu_1 - \widehat{x})$ or equivalently, if $0 < \sigma^2 \leq (1 - \mu_1)(\mu_1 - \widehat{x})$
2. For $\mu_1 < \widehat{x}$,
   - The probability distribution does not exist, if $\mu_1^2 < \mu_2 \leq \mu_1^2 + \mu_1(\widehat{x} - \mu_1)$ or equivalently, if $0 < \sigma^2 \leq \mu_1(\widehat{x} - \mu_1)$
3. For $\mu_1 = \widehat{x}$,
   - The probability distribution exists anyway, because the condition (11.52) simplifies itself to $\mu_2 > \mu_1^2$

11.10 Continuous uni-extremal probability distribution

11.10.1 The general case

For a continuous $Y|\{d_Y\}$, such that $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$ or equivalently $d = (\mu_1, \mu_2)$, then the probability density functions of $Y|\{d_Y\}$ and $X|\{d\}$ are

$$\begin{aligned} f_{Y|\{d_Y\}}(y) &= \frac{e^{\lambda_1 y+\lambda_2 y^2}}{\int_a^b e^{\lambda_1 y+\lambda_2 y^2} dy},\ y \in \mathcal{X}_Y(\{d_Y\}) = [a,b] \\ \text{and } f_{X|\{d\}}(x) &= \frac{e^{\beta x+\gamma x^2}}{\int_0^1 e^{\beta x+\gamma x^2} dx},\ x \in \mathcal{X}_X(\{d\}) = [0,1] \end{aligned} \tag{11.53}$$

respectively.

In order to perform this, the following system of three simultaneous equations has to be solved for $\lambda(d_Y) = (\lambda_1, \lambda_2)$:

$$\int_a^b e^{\lambda_0+\lambda_1 y+\lambda_2 y^2} dy = 1 \tag{11.54}$$

$$\int_a^b y e^{\lambda_0+\lambda_1 y+\lambda_2 y^2} dy = \mu_Y^{(1)} \tag{11.55}$$

$$\int_a^b y^2 e^{\lambda_0+\lambda_1 y+\lambda_2 y^2} dy = \mu_Y^{(2)} \tag{11.56}$$

which is equivalent to the solution of the following system of simultaneous equations in λ_1 and λ_2:

$$\begin{cases} \mu_Y^{(1)} &= \dfrac{\int_a^b y e^{\lambda_1 y+\lambda_2 y^2} dy}{\int_a^b e^{\lambda_1 y+\lambda_2 y^2} dy} \\ \mu_Y^{(2)} &= \dfrac{\int_a^b y^2 e^{\lambda_1 y+\lambda_2 y^2} dy}{\int_a^b e^{\lambda_1 y+\lambda_2 y^2} dy} \end{cases} \tag{11.57}$$

Therefore, in order to compute the desired $\lambda(d_Y) = (\lambda_1, \lambda_2)$, the solution of (11.57) is yielded by the solving of the following simultaneous system of equations in β and γ at first:

$$\begin{cases} \mu_1 &= \frac{\int\limits_0^1 x e^{\beta x+\gamma x^2} dx}{\int\limits_0^1 e^{\beta x+\gamma x^2} dx} \\ \mu_2 &= \frac{\int\limits_0^1 x^2 e^{\beta x+\gamma x^2} dx}{\int\limits_0^1 e^{\beta x+\gamma x^2} dx} \end{cases} \tag{11.58}$$

and only thereafter we arrive at

- $\lambda_1 = \frac{\beta(b-a)-2a\gamma}{(b-a)^2}$ and
- $\lambda_2 = \frac{\gamma}{(b-a)^2}$
- $\lambda_0 = -\log\left(\int\limits_a^b e^{\lambda_1 y+\lambda_2 y^2} dy\right)$

  For the purpose of handling the overflow computing errors, λ_0 is computed according to the following rule:

  - if $(\beta + \gamma < 709)$ then $\lambda_0 = -\frac{\beta a}{b-a} + \frac{\gamma a^2}{(b-a)^2} - \log\left(\int\limits_0^1 e^{\beta x+\gamma x^2} dx\right)$
  - if $(\beta + \gamma \geq 709)$ then $\lambda_0 = -\beta - \gamma - \frac{\beta a}{b-a} + \frac{\gamma a^2}{(b-a)^2} - \log\left(\int\limits_0^1 e^{\beta(x-1)+\gamma(x^2-1)} dx\right)$

11.10.2 The cases for symmetry and constancy

This is the case, when $\mu_Y^{(1)} = \frac{a+b}{2}$ or equivalently $\mu_1 = \frac{1}{2}$. Subsequently, $\beta = -\gamma$.

Here, the probability distribution is symmetric but not constant, if $\mu_1 = \frac{1}{2}$ and $\mu_2 \neq \frac{1}{3}$. In that case,

- $\lambda_1 = \frac{\beta}{b-a} + \frac{2a\beta}{(b-a)^2}$
- $\lambda_2 = -\frac{\beta}{(b-a)^2}$

- $\lambda_0 = -\log(b-a) - \log\left(\int_0^1 e^{\beta x - \beta x^2} dx\right) - \frac{\beta a}{b-a} - \frac{\beta a^2}{(b-a)^2}$

Again, the probability distribution is constant (namely uniform), if $\mu_1 = \frac{1}{2}$ and $\mu_2 = \frac{1}{3}$. In this case, $\beta = \gamma = 0$, so that

- $\lambda_1 = 0$
- $\lambda_2 = 0$
- $\lambda_0 = -\log(b-a)$

11.10.3 The monotonic case

Let β_0 be the (unique) solution of the following equation in β:

$$\mu_1 = \frac{\int_0^1 x e^{\beta x} dx}{\int_0^1 e^{\beta x} dx} = 1 + \frac{1}{e^\beta - 1} - \frac{1}{\beta} \text{ (provided } \beta \neq 0)$$

Then, if so happens that β_0 fulfills the following equality:

$$\mu_2 = \frac{\int_0^1 x^2 e^{\beta_0 x} dx}{\int_0^1 e^{\beta_0 x} dx} = 1 + \frac{1}{e^\beta - 1} - \frac{2\mu_1}{\beta} \text{ (provided } \beta_0 \neq 0)$$

then the probability distribution is a monotonic probability distribution, which means

- $\lambda_0 = -\log(b-a) - \log\left(\int_0^1 e^{\beta_0 x} dx\right) - \beta_0 \left(\frac{a}{b-a}\right)$
- $\lambda_1 = \frac{\beta_0}{b-a}$
- $\lambda_2 = 0$

Notably, if $\beta_0 = 0$ then the probability distribution is uniform.

Otherwise, if we are sure that $\beta_0 \neq 0$, then we can easily rewrite $\lambda_0 = \log\left(\frac{\beta_0}{(b-a)(e^{\beta_0}-1)}\right) - \beta_0 \left(\frac{a}{b-a}\right), \lambda_1 = \frac{\beta_0}{b-a}$ and $\lambda_2 = 0$.

11.10.4 Uni-extremal probability distribution with a small variance

Let μ_Y and σ_Y^2 be the respective **user-given** mean and variance of the uni-extremal probability distribution of Y with the given compact support $[a, b]$.

Let the random variable Y be assumed to follow a **truncated normal distribution** with the compact support $[a, b]$, whose probability density function $f_{Y|\{\widetilde{d}_Y\}}(y)$, $a \leq y \leq b$ determined by $\widetilde{d}_Y = (\widetilde{\mu}_Y^{(1)}, \widetilde{\mu}_Y^{(2)})$, such that $\widetilde{\sigma}_Y^2 = \widetilde{\mu}_Y^{(2)} - (\widetilde{\mu}_Y^{(1)})^2$, is given by

$$\begin{aligned} f_{Y|\{\widetilde{d}_Y\}}(y) &= \frac{e^{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2}}{(b-a)\int\limits_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx}, \quad a \leq y \leq b \\ &= \frac{K}{\sigma_Y\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2}, \quad a \leq y \leq b \end{aligned} \tag{11.59}$$

such that $\mu_1 = \frac{\mu_Y - a}{b-a}$ and $\sigma = \frac{\sigma_Y}{b-a}$.

By the linear transformation $X = \frac{Y-a}{b-a}$, the probability density function $f_{X|\{\widetilde{d}\}}(x)$, $0 \leq x \leq 1$ of the random variable X is given by

$$f_{X|\{\widetilde{d}\}}(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2}}{\int\limits_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx}, \quad 0 \leq x \leq 1 \tag{11.60}$$

such that

- $K = \frac{\sigma_Y\sqrt{2\pi}}{(b-a)\int\limits_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx}$ and $e^{-\widetilde{\alpha}} = \int\limits_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx$
- for $x = \frac{y-a}{b-a}$, $f_{X|\{\widetilde{d}\}}(x) = (b-a) f_{Y|\{\widetilde{d}_Y\}}(y)$

We shall make use of the values of μ_1 and σ to judge, whether the uni-extremal probability distribution (11.59) of Y is acceptable as the approximating probability distribution of Y with small variance (having mean $\widetilde{\mu}_Y$ and standard deviation $\widetilde{\sigma}_Y$, such that $\widetilde{\mu}_Y \approx \mu_Y$ and $\widetilde{\sigma}_Y \approx \mu_Y$).

Therefore, let us **derive** and **examine** the **smallness** of the values of $\epsilon_{\mu_1} = \widetilde{\mu}_1 - \mu_1$ and $\epsilon_\sigma = \widetilde{\sigma} - \sigma$, where $\widetilde{\mu}_1$ and $\widetilde{\sigma}$ being the mean and the standard deviation of the probability distribution of X given by (11.60) respectively.

At first, ϵ_{μ_1} is derived as

$$\begin{aligned}
\epsilon_{\mu_1} &= e^{\widetilde{\alpha}} \int_0^1 x e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx - \mu_1 \text{ by putting } \widetilde{\alpha} = -\log\left(\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx\right) \\
&= e^{\widetilde{\alpha}} \int_0^1 (x-\mu_1) e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx \\
&= e^{\widetilde{\alpha}} \left(\sigma\sqrt{2}\right)^2 \int_0^1 \left(\frac{x-\mu_1}{\sigma\sqrt{2}}\right) e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} \frac{1}{\sigma\sqrt{2}}\ dx \\
&= e^{\widetilde{\alpha}} \sigma^2 \left[-e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2}\Big|_{x=0}^{x=1}\right] \\
&= e^{\widetilde{\alpha}} \sigma^2 \left[e^{-\frac{\mu_1^2}{2\sigma^2}} - e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}\right] \\
&= \sigma^2 \frac{e^{-\frac{\mu_1^2}{2\sigma^2}} - e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx} \\
&= \sigma \frac{e^{-\frac{\mu_1^2}{2\sigma^2}} - e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\int_{-\frac{\mu_1}{\sigma}}^{\frac{1-\mu_1}{\sigma}} e^{-\frac{x^2}{2}} dx} = \sigma \frac{e^{-\frac{\mu_1^2}{2\sigma^2}} - e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\sqrt{2\pi}\frac{1}{\sqrt{2\pi}}\int_{-\frac{\mu_1}{\sigma}}^{\frac{1-\mu_1}{\sigma}} e^{-\frac{x^2}{2}} dx} \\
&= \frac{\sigma}{\sqrt{2\pi}} \frac{e^{-\frac{\mu_1^2}{2\sigma^2}} - e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(\frac{-\mu_1}{\sigma}\right)} \qquad (11.61)
\end{aligned}$$

In addition to this, we shall need the following result

$$\frac{\epsilon_{\mu_1}}{\sigma} = \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{\mu_1^2}{2\sigma^2}} - e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(\frac{-\mu_1}{\sigma}\right)} \qquad (11.62)$$

As the next step, by making use of the following **indefinite integral**

$$\int (x-\mu_1)\, e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx = -\sigma^2 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} \tag{11.63}$$

on integrating by parts, we get

$$\int (x-\mu_1)^2\, e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx = -(x-\mu_1)\sigma^2 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} + \sigma^2 \int e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx \tag{11.64}$$

which leads us to

$$\begin{aligned} \int_0^1 (x-\mu_1)^2 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx &= \sigma^2 \left\{ \int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx - (1-\mu_1) e^{-\frac{1}{2}\left(\frac{1-\mu_1}{\sigma}\right)^2} - \mu_1 e^{-\frac{1}{2}\left(\frac{\mu_1}{\sigma}\right)^2} \right\} \\ &= \sigma^2 \left(e^{-\widetilde{\alpha}} - \mu_1 e^{-\frac{1}{2}\left(\frac{\mu_1}{\sigma}\right)^2} - (1-\mu_1) e^{-\frac{1}{2}\left(\frac{1-\mu_1}{\sigma}\right)^2} \right) \end{aligned} \tag{11.65}$$

Therefore, by using (11.65) we shall derive the variance $\widetilde{\sigma}^2$ of X as the next step as follows:

$$\begin{aligned} \widetilde{\sigma}^2 &= E[(X-\widetilde{\mu}_1)^2] = E[(X-\mu_1-\epsilon_{\mu_1})^2] = E[(X-\mu_1)^2] - \epsilon_{\mu_1}^2 \\ &= e^{\widetilde{\alpha}} \int_0^1 (x-\mu_1)^2\, e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx - \epsilon_{\mu_1}^2 \\ &= e^{\widetilde{\alpha}} \sigma^2 \left(e^{-\widetilde{\alpha}} - \mu_1 e^{-\frac{1}{2}\left(\frac{\mu_1}{\sigma}\right)^2} - (1-\mu_1) e^{-\frac{1}{2}\left(\frac{1-\mu_1}{\sigma}\right)^2} \right) - \epsilon_{\mu_1}^2 \\ &= \sigma^2 \left[1 - \frac{(1-\mu_1) e^{-\frac{1}{2}\left(\frac{1-\mu_1}{\sigma}\right)^2} + \mu_1 e^{-\frac{1}{2}\left(\frac{\mu_1}{\sigma}\right)^2}}{\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx} \right] - \epsilon_{\mu_1}^2 \\ &= \sigma^2 - \frac{\sigma \left[(1-\mu_1) e^{-\frac{1}{2}\left(\frac{1-\mu_1}{\sigma}\right)^2} + \mu_1 e^{-\frac{1}{2}\frac{\mu_1^2}{\sigma^2}} \right]}{\sqrt{2\pi}\frac{1}{\sqrt{2\pi}} \int_{-\frac{\mu_1}{\sigma}}^{\frac{1-\mu_1}{\sigma}} e^{-\frac{x^2}{2}} dx} - \epsilon_{\mu_1}^2 \\ &= \sigma^2 - \frac{\sigma}{\sqrt{2\pi}} \frac{\mu_1 e^{-\frac{\mu_1^2}{2\sigma^2}} + (1-\mu_1) e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(-\frac{\mu_1}{\sigma}\right)} - \epsilon_{\mu_1}^2 \end{aligned} \tag{11.66}$$

which gives

$$\widetilde{\sigma}^2 - \sigma^2 = -\epsilon_{\mu_1}^2 - \frac{\sigma}{\sqrt{2\pi}} \frac{\mu_1 e^{-\frac{\mu_1^2}{2\sigma^2}} + (1-\mu_1) e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(\frac{-\mu_1}{\sigma}\right)} \tag{11.67}$$

In the final step, ϵ_σ can be derived by using (11.66) and (11.62) as follows

$$\begin{aligned}
\epsilon_\sigma &= \widetilde{\sigma} - \sigma \\
&= \left\{ \sigma^2 - \frac{\sigma}{\sqrt{2\pi}} \frac{\mu_1 e^{-\frac{\mu_1^2}{2\sigma^2}} + (1-\mu_1) e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(-\frac{\mu_1}{\sigma}\right)} - \epsilon_{\mu_1}^2 \right\}^{\frac{1}{2}} - \sigma \\
&= \sigma \left\{ 1 - \frac{1}{\sqrt{2\pi}} \frac{\left(\frac{\mu_1}{\sigma}\right) e^{-\frac{\mu_1^2}{2\sigma^2}} + \left(\frac{1-\mu_1}{\sigma}\right) e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(-\frac{\mu_1}{\sigma}\right)} - \left(\frac{\epsilon_{\mu_1}}{\sigma}\right)^2 \right\}^{\frac{1}{2}} - \sigma \\
&= \sigma \left\{ 1 - \frac{1}{\sqrt{2\pi}} \frac{\left(\frac{\mu_1}{\sigma}\right) e^{-\frac{\mu_1^2}{2\sigma^2}} + \left(\frac{1-\mu_1}{\sigma}\right) e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(-\frac{\mu_1}{\sigma}\right)} \right. \\
&\qquad \left. - \frac{1}{2\pi} \left(\frac{e^{-\frac{\mu_1^2}{2\sigma^2}} - e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(-\frac{\mu_1}{\sigma}\right)} \right)^2 \right\}^{\frac{1}{2}} - \sigma \qquad (11.68) \\
&= \sigma \left\{ -\frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{\left(\frac{\mu_1}{\sigma}\right) e^{-\frac{\mu_1^2}{2\sigma^2}} + \left(\frac{1-\mu_1}{\sigma}\right) e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(-\frac{\mu_1}{\sigma}\right)} \right. \\
&\qquad \left. - \frac{1}{4\pi} \left(\frac{e^{-\frac{\mu_1^2}{2\sigma^2}} - e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(-\frac{\mu_1}{\sigma}\right)} \right)^2 + O\left(\left(\frac{\theta}{\sigma}\right)^2 e^{-\left(\frac{\theta}{\sigma}\right)^2} \right) \right\}, 0 < \theta < 1 \\
&= -\frac{1}{2\sqrt{2\pi}} \frac{\mu_1 e^{-\frac{\mu_1^2}{2\sigma^2}} + (1-\mu_1) e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(-\frac{\mu_1}{\sigma}\right)} - \frac{\sigma}{4\pi} \left(\frac{e^{-\frac{\mu_1^2}{2\sigma^2}} - e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}}{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(-\frac{\mu_1}{\sigma}\right)} \right)^2 \\
&\quad + O\left(\widetilde{\ell}^2\right) \qquad (11.69)
\end{aligned}$$

where $\widetilde{\ell} = \left(\frac{\theta}{\sigma}\right) e^{-\frac{1}{2}\left(\frac{\theta}{\sigma}\right)^2}$ in this case, such that $\theta \in \{\mu_1, 1-\mu_1\}$ for $0 < \mu_1 < 1$ and the remainder $O\left(\widetilde{\ell}^2\right)$ gives the terms containing powers of $\widetilde{\ell}$ **higher** than **one**.

Therefore, for **smaller** values of σ, it can be easily seen, that

- $\frac{\mu_1^2}{\sigma^2}$ and $\frac{(1-\mu_1)^2}{\sigma^2}$ have **larger** values and therefore $e^{-\frac{\mu_1^2}{2\sigma^2}}$ or $e^{-\frac{(1-\mu_1)^2}{2\sigma^2}}$ have **extreme smaller** values
- The values of $\widetilde{\ell}$ tend to be **smaller** with the **decrease** in σ and in fact, $\lim_{\sigma \to 0} \left(\frac{\theta}{\sigma}\right) e^{-\frac{1}{2}\left(\frac{\theta}{\sigma}\right)^2} = 0$
- $\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(\frac{-\mu_1}{\sigma}\right) \approx 1$ and in fact, $\lim_{\sigma \to 0} \left\{\Phi\left(\frac{1-\mu_1}{\sigma}\right) - \Phi\left(\frac{-\mu_1}{\sigma}\right)\right\} = 1$

This leads us to **justify** the **usage** of the **truncated normal probability distribution** as a good approximation to the desired continuous approximated minimum information uni-extremal probability distribution. In other words, the reason for this justification is nothing different from the **smallness** of the values of ϵ_{μ_1} and ϵ_σ, the detailed description of which are given in (11.61) and (11.69) (or in (11.68) for the sake of exactness) respectively.

However, for the sake of **programming simplicity**, my **java language code** meant for the computation of ϵ_σ uses the following formula:

$$\epsilon_\sigma = \widetilde{\sigma} - \sigma = \sqrt{\epsilon_{\sigma^2} + \sigma^2} - \sigma \tag{11.70}$$

such that $\epsilon_\sigma^2 = \widetilde{\sigma}^2 - \sigma^2$ is given by (11.67).

As we have already discussed before, the approximating condition, for which (11.60) is accepted to be the probability distribution of X with subject to the user-wished mean (μ_1) and the standard deviation (σ), reads $\sigma < \frac{\mu_D}{\ell}$, where $\mu_D = \min(\mu_1, 1 - \mu_1)$.

Obviously, a particular extent of the smallness of the expressions ϵ_{μ_1} and ϵ_σ contribute to the fulfillment of the aforesaid approximating condition.

Consequently, this approximating condition is also valid for assuming (11.59) to be the probability distribution of Y, with subject to the user-wished mean (μ_Y) and the standard deviation (σ_Y).

Therefore, we can come to the following conclusion:

(11.59) is assumed to be the approximated probability distribution of Y, if the following case arises

- $\sigma < \frac{\mu_D}{\ell}$. In our cases, ℓ is made to range from 2.08699 to 3.5

Now, let us come to the determination of λ values. They read as follows:

- $\lambda_0 = \log\left(\frac{K}{\sigma_Y\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{\mu_Y}{\sigma_Y}\right)^2$
- $\lambda_1 = \frac{\mu_Y}{\sigma_Y^2}$
- $\lambda_2 = -\frac{1}{2\sigma_Y^2}$

11.11 The overflow and underflow errors

The software programs, which are to be developed so as to fulfill the **main target** of this dissertation, are developed in the object oriented programming language named Java.

Like any program executer, the Java executer does not admit any arbitrarily **large** number or any arbitrarily **small** number **in terms of magnitude**.

Therefore, we enlist the maximum and the minimum allowable real numbers (in terms of their magnitudes) with their natural logarithmic values as follows:

- The **maximum** allowable number reads $Double.MAX\_VALUE = 1.7976931348623157 \times 10^{308}$, whose natural logarithm is $\log(Double.MAX\_VALUE) = 709.782712893384$.

  So, if a computed number **exceeds** $Double.MAX\_VALUE$, then would be an **overflow error**.

- The **minimum** allowable number reads $Double.MIN\_VALUE = 4.9 \times 10^{-324}$, whose natural logarithm is $\log(Double.MAX\_VALUE) = -744.4400719213812$.

  So, if a computed number **falls below** $Double.MIN\_VALUE$, then there would be an **underflow error**.

In course of my programming work, I had to take care to handle and resolve the problems, in the cases when the exponential power of e **exceeded** 709 or **fell below** -744.

Such problems were **well resolvable** by using certain **mathematical tricks**.

Chapter 12

Numerical algorithms

In order to determine the minimum information probability distribution, where the numerical values of the first m moments are given, the coefficients $\lambda_0, \lambda_1, \ldots, \lambda_m$ are to be computed numerically by solving a System of $(m+1)$ simultaneous equations. As a matter of fact, a solution to every system of equations corresponding to an arbitrary set of given values of the moments $\mu_1, \ldots, \mu_n$ does not exist, because the moments have definite regions of validity and cannot assume arbitrary real values. These regions of validity must be determined for each given value of m. Therefore, individual considerations for different values of m are absolutely necessary.

As we have already discussed, our numerical algorithms shall be confined to $m \leq 2$. Moreover, in cases for $m \leq 2$, a minimum information probability distribution is based on a maximum entropy probability distribution (MEP-probability distribution) for $m \leq 2$ moments. However, no numerical treatments are necessary in trivial cases for $m = 0$.

The determination of a minimum information probability distribution demands a numerical solution to an equation or the same of a system of simultaneous equations. This numerical solution is the heart of the problem. So far, the Newton Raphson procedure is the most ideal procedure for the numerical solutions of our problems. However, the Newton Raphson procedure demands a predetermined approximated solution, without which the procedure generally turns out to be a failure. Keeping this in view, the numerical solutions are discussed in the coming two sections.

12.1 Numerical integration by Weddle's rule

12.1.1 The plan of action

This dissertation handles the numerical integrations of certain continuous functions only. The **Weddle's rule of the numerical integration**, which belongs to one of the **Newton-Cotes-Formulae** (these formulae are well stated in the page 116 of [40]), happens to be one of the best numerical integration procedures for handling these kinds of continuous integrands, both in terms of speed and accuracy.

In a plain and simple language, the Newton-Cotes-formula with subject to $\mathbf{n} = 6$ is the aforesaid Weddle's rule, where **n** is the degree of **the interpolating polynomial** $f_0(x)$ used for deriving the preliminary Weddle's integral (12.5), namely the integral $\int_{x_0}^{x_6} f_0(x)dx$, such that the difference $x_{i+1} - x_i$ is constant for every $i \in \{0, 1, 2, 3, 4, 5\}$. After this, with subject to a suitable choice of a natural number n as a multiple of 6, these Weddle's integrals of the form $\int_{x_{6i}}^{x_{6(i+1)}} f_0(x)dx$, $i \in \{0, 1, 2, \ldots, \frac{n-6}{6}\}$ are therefore summed up together to give the final Weddle's rule of numerical integration (12.6), namely the integral $\int_{x_0}^{x_n} f_0(x)dx$ for our programming work. The technique of how to choose this particular n shall be discussed in due course.

This interpolating polynomial $f_0(x)$ of degree 6 is calculated by means of the well known **Newton's forward interpolation formula** (referred to the pages 56 - 58 of [36]).

A reader of this dissertation may put a very logical question: Why should the degree of the interpolating polynomial be limited to 6? In accordance with a clear statement given in the page 116 of [40], the answer is, any value of **n** higher than 6 would make the coefficients of the Weddle's formula negative and thereby making the Weddle's formula completely useless.

Before we proceed to discuss the Weddle's numerical integration procedure in the full details, we wish to introduce our readers to the **forward difference operator $\boldsymbol{\Delta}$** and the **shift operator** E (these operators are well explained in the pages 33 - 37 of [35]). These operators and their related useful results

shall be used for our coming derivations and hence their introductions are absolutely needed for the sake of clarity.

Let a continuous function $f(x)$ be defined in a closed interval $[a, b]$, which is divided into n ($n \in \mathbb{N}^+$) sub-intervals $[x_i, x_{i+1}]$, ($i = 0, 1, \ldots, n$) of equal length h, where $a = x_0,\ b = x_n$ and $h = x_{i+1} - x_i$ $(i = 0, 1, \ldots, n)$.

In that case, there are exactly $(n + 1)$ equidistant arguments x_i and $(n + 1)$ entries $y_i = f(x_i)$ ($i = 0, 1, \ldots, n$).

At first we shall introduce two important operators, namely the forward difference operator $\boldsymbol{\Delta}$ and the shift operator E :

- $\boldsymbol{\Delta}$ is defined by

$$\boldsymbol{\Delta} y_i = y_{i+1} - y_i = f(x_{i+1}) - f(x_i) = f(x_i + h) - f(x_i)$$

- and E is defined by

$$E y_i = y_{i+1} = f(x_{i+1}) = f(x_i + h)$$

Next we shall introduce four important results:

1.

$$E \equiv 1 + \boldsymbol{\Delta},\ \text{since}\ E y_i = y_{i+1} = y_{i+1} - y_i + y_i = \boldsymbol{\Delta} y_i + y_i = (1 + \boldsymbol{\Delta}) y_i$$

2.

$$\begin{aligned}
\boldsymbol{\Delta}^2 y_i &= \boldsymbol{\Delta}(\boldsymbol{\Delta} y_i) = \boldsymbol{\Delta}(y_{i+1} - y_i) = (y_{i+2} - y_{i+1}) - (y_{i+1} - y_i) \\
&= y_{i+2} - 2y_{i+1} + y_i \\
\Rightarrow \boldsymbol{\Delta}^3 y_i &= \boldsymbol{\Delta}(\boldsymbol{\Delta}^2 y_i) = \boldsymbol{\Delta}(y_{i+2} - 2y_{i+1} + y_i) \\
&= (y_{i+3} - y_{i+2}) - 2(y_{i+2} - y_{i+1}) + (y_{i+1} - y_i) \\
&= y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i \\
&\vdots \\
\Rightarrow \boldsymbol{\Delta}^p y_i &= y_{p+i} - p y_{p-1+i} + \binom{p}{2} y_{p-2+i} - \binom{p}{3} y_{p-3+i} + \cdots + (-1)^p y_i
\end{aligned}$$

for any $p \in \mathbb{N}^+$.

3.
$$\begin{aligned} E^2 y_i &= E(Ey_i) = E(y_{i+1}) = y_{i+2} \\ \Rightarrow E^3 y_i &= E(E^2 y_i) = E(y_{i+2}) = y_{i+3} \\ &\vdots \\ \Rightarrow E^p y_i &= y_{p+i} \ , \ (p \in \mathbb{N}^+) \end{aligned}$$

4. Under the assumption that the derivatives of $f(x)$ with respect to x exist at least upto the p^{th} order,

$$\begin{aligned} \boldsymbol{\Delta}^p y_i &= \boldsymbol{\Delta}^{p-1}(y_{i+1} - y_i) = \boldsymbol{\Delta}^{p-1}(f(x_i + h) - f(x_i)) \\ &= \boldsymbol{\Delta}^{p-1} h f^{'}(x_i + \theta_1 h) \ , \ 0 < \theta_1 < 1 \\ \\ &= \boldsymbol{\Delta}^{p-2} h (f^{'}(x_i + h + \theta_1 h) - f^{'}(x_i + \theta_1 h)) \\ &= \boldsymbol{\Delta}^{p-2} h^2 f^{''}(x_i + \theta_1 h + \theta_2 h) \ , \ 0 < \theta_2 < 1 \\ &= \boldsymbol{\Delta}^{p-2} h^2 f^{''}(x_i + (\theta_1 + \theta_2) h) \\ \\ &= \boldsymbol{\Delta}^{p-3} h^3 f^{'''}(x_i + (\theta_1 + \theta_2 + \theta_3) h) \ , \ 0 < \theta_3 < 1 \\ &\vdots \\ &= h^p f^{(p)}(x_i + (\theta_1 + \theta_2 + \ldots + \theta_p) h) \ , \ 0 < \theta_j < 1 \ , j = 1, 2, \ldots, p \\ &= O(h^p) \end{aligned}$$

However, it has to be clearly stated that the usage of the ordered notation, namely $O(h^p)$, happens to be meaningful, only when $0 < h < 1$, otherwise meaningless in this regard.

So, for the sake of the best degree of this dissertation's clarity, the Newton's forward interpolation formula (referred to the pages 56 to 58 of [36]) is derived at first and subsequently the Weddle's rule of numerical integration is derived.

The derivation of the estimation of error in the Weddle's integral (the statement of this derived estimation can also be refereed to the page 116 of [40]) has included thereafter. The **smallness of this estimation of error** shows the justification of the usage of the Weddle's rule of numerical integration for our programming works.

Therefore, by keeping these points in mind, we shall proceed to derive the Newton's forward interpolation formula by **finite mathematical induction** formally and rigorously, as already mentioned. This formula has been briefly outlined in the pages 56 - 58 of [36], but without being proved rigorously.

This interpolation formula refers to the evaluation of the function $f(x)$ at a given value x, where x is any value lying between x_0 and x_n (i.e. $x \in [x_0, x_n]$), completely disregarding the form of f, but with the help of the given equidistant arguments x_i and the corresponding given entries $y_i = f(x_i),\ i = 0, 1, \ldots, n$.

12.1.2 Newton's forward interpolation formula:

Since the values of the function $f(x)$ are known at the equidistant points x_i $(i = 0, 1, \ldots, n)$, $f(x)$ can be approximated to a polynomial $f_0(x)$ of degree n as $f(x) \approx f_0(x)$, such that

$$\begin{aligned} f(x_i) &= f_0(x_i) \text{ for } i = 0, 1, \ldots, n \\ f(x) &\approx f_0(x) \text{ for } x \in [x_0, x_n] \end{aligned} \tag{12.1}$$

$f_0(x)$ is said to be the interpolating polynomial of $f(x)$, which is given by

$$f_0(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}) \tag{12.2}$$

which means, $f_0(x)$ is thus an expression with $(n+1)$ arguments and $(n+1)$ unknowns a_i $(i = 0, 1, \ldots, n)$. These a_i can be calculated as follows:

$$y_0 = f_0(x_0) = a_0$$

$$y_1 = f_0(x_1) = a_0 + a_1(x_1 - x_0) = y_0 + ha_1 \Rightarrow a_1 = \frac{\mathbf{\Delta} y_0}{h}$$

$$y_2 = f_0(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)$$

$$= y_0 + 2h\frac{\mathbf{\Delta} y_0}{h} + a_2(2h)(h) = y_0 + 2\mathbf{\Delta} y_0 + 2h^2 a_2 = y_0 + 2(y_1 - y_0) + 2h^2 a_2$$

$$\Rightarrow 2h^2 a_2 = y_2 - 2y_1 + y_0 = \mathbf{\Delta}^2 y_0 \Rightarrow a_2 = \frac{\mathbf{\Delta}^2 y_0}{2h^2}$$

At this point we shall assert $a_i = \dfrac{\mathbf{\Delta}^i y_0}{h^i\, i!}$, $i = 0, 1, \ldots, n$ which needs be proved by finite mathematical induction. We have already shown that the result is valid for $i = 0, 1, 2$. Thus we have only to show the induction step $i \to i + 1$, which means that assuming that the result is valid for i, we need to prove the validity of the result for $i + 1$:

The induction step $i \to i+1$:

$$\begin{aligned}
y_{i+1} = f_0(x_{i+1}) &= y_0 + (x_{i+1}-x_0)\frac{\boldsymbol{\Delta} y_0}{h} + (x_{i+1}-x_0)(x_{i+1}-x_1)\frac{\boldsymbol{\Delta}^2 y_0}{h^2 2!} \\
&\quad + \ldots + (x_{i+1}-x_0)(x_{i+1}-x_1)\ldots(x_{i+1}-x_{i-1})\frac{\boldsymbol{\Delta}^i y_0}{h^i i!} \\
&\quad + a_{i+1}(x_{i+1}-x_0)(x_{i+1}-x_1)\ldots(x_{i+1}-x_i) \\
&= y_0 + (i+1)h\frac{\boldsymbol{\Delta} y_0}{h} + (i+1)(i)h^2\frac{\boldsymbol{\Delta}^2 y_0}{h^2 2!} + \ldots + (i+1)(i)\ldots 2\, h^i \frac{\boldsymbol{\Delta}^i y_0}{h^i i!} \\
&\quad + a_{i+1}(i+1)(i)\ldots 1\, h^{i+1} \\
&= y_0 + \binom{i+1}{1}\boldsymbol{\Delta} y_0 + \binom{i+1}{2}\boldsymbol{\Delta}^2 y_0 + \ldots + \binom{i+1}{i}\boldsymbol{\Delta}^i y_0 + a_{i+1}h^{i+1}(i+1)! \\
&= y_0 + \binom{i+1}{1}\boldsymbol{\Delta} y_0 + \binom{i+1}{2}\boldsymbol{\Delta}^2 y_0 + \ldots + \binom{i+1}{i+1}\boldsymbol{\Delta}^{i+1} y_0 - \binom{i+1}{i+1}\boldsymbol{\Delta}^{i+1} y_0 \\
&\quad + a_{i+1}h^{i+1}(i+1)! \\
&= \left[1 + \binom{i+1}{1}\boldsymbol{\Delta} + \binom{i+1}{2}\boldsymbol{\Delta}^2 + \ldots + \binom{i+1}{i+1}\boldsymbol{\Delta}^{i+1}\right] y_0 - \boldsymbol{\Delta}^{i+1} y_0 \\
&\quad + a_{i+1}h^{i+1}(i+1)! \\
&= (1+\boldsymbol{\Delta})^{i+1} y_0 - \boldsymbol{\Delta}^{i+1} y_0 + a_{i+1}h^{i+1}(i+1)! \\
&= E^{i+1} y_0 - \boldsymbol{\Delta}^{i+1} y_0 + a_{i+1}h^{i+1}(i+1)! \\
&= y_{i+1} - \boldsymbol{\Delta}^{i+1} y_0 + a_{i+1}h^{i+1}(i+1)! \\
\iff a_{i+1} &= \frac{\boldsymbol{\Delta}^{i+1} y_0}{h^{i+1}(i+1)!}
\end{aligned}$$

□

Hence, the proved result $a_i = \dfrac{\boldsymbol{\Delta}^i y_0}{h^i\ i!}, i = 0, 1, \ldots, n$ are the derived coefficients

of the constructed polynomial (12.2) and thus the interpolating polynomial $f_0(x)$ of degree n with known coefficients is given as follows:

$$y = f(x) \approx f_0(x) = y_0 + (x - x_0)\frac{\mathbf{\Delta} y_0}{h} + (x - x_0)(x - x_1)\frac{\mathbf{\Delta}^2 y_0}{h^2 2!} + \ldots + (x - x_0)(x - x_1)\ldots(x - x_{n-1})\frac{\mathbf{\Delta}^n y_0}{h^n n!} \tag{12.3}$$

which is the **Newton's forward interpolation formula**.

Now, in the final step, we shall derive **Weddle's rule of numerical integration** using Newton's forward interpolation formula:

12.1.3 Weddle's rule of numerical integration

Let $[c, d]$ be a closed interval, where $f(x)$ is continuous and is approximated by a polynomial of degree $n = 6$ by using Newton's forward interpolation formula as:

$$y \approx f_0(x) = y_0 + (x - x_0)\frac{\mathbf{\Delta} y_0}{h} + (x - x_0)(x - x_1)\frac{\mathbf{\Delta}^2 y_0}{h^2 2!} + \ldots + (x - x_0)(x - x_1)\ldots(x - x_5)\frac{\mathbf{\Delta}^6 y_0}{h^6 6!}$$

so that, $x_0 = c$, $x_6 = d = c + 6h$ and $x_0, x_1, \ldots, x_6, y_0, y_1, \ldots, y_6$ are all known.

Taking $x = x_0 + sh$, we get the interpolating polynomial $f_0(x)$ of degree 6 as

$$y \approx f_0(x) = y_0 + s\mathbf{\Delta} y_0 + s(s-1)\frac{\mathbf{\Delta}^2 y_0}{2!} + \ldots + s(s-1)\ldots(s-5)\frac{\mathbf{\Delta}^6 y_0}{6!}$$

Here,

$$\int_c^d f_0(x)\,dx = \int_{x_0}^{x_6} f_0(x)\,dx = \int_{s=0}^{s=6} f_0(x)(h\,ds) = h\int_0^6 f_0(x)\,ds$$

$$= h\int_0^6 \left[y_0 + s\mathbf{\Delta} y_0 + s(s-1)\frac{\mathbf{\Delta}^2 y_0}{2!} + \ldots + s(s-1)\ldots(s-5)\frac{\mathbf{\Delta}^6 y_0}{6!}\right] ds$$

Now, this integral has to be evaluated part by part. It is to be noted that, if the original function $f(x)$ is a polynomial of degree at most 6, then the numerical integral would give the exact results. We have:

$$\int_0^6 s\,ds = \frac{6^2}{2} = 18$$

$$\int_0^6 s(s-1)\,ds = \frac{6^3}{3} - \frac{6^2}{2} = 54$$

$$\int_0^6 s(s-1)(s-2)\,ds = \int_0^6 (s^3 - 3s^2 + 2s)\,ds = \frac{6^4}{4} - \frac{3.6^3}{3} + 6^2 = 144$$

$$\int_0^6 s(s-1)(s-2)(s-3)\,ds = \int_0^6 (s^4 - 6s^3 + 11s^2 - 6s)\,ds$$

$$= \frac{6^5}{5} - \frac{6.6^4}{4} + \frac{11.6^3}{3} - \frac{6.6^2}{2} = \frac{1476}{5}$$

$$\int_0^6 s(s-1)(s-2)(s-3)(s-4)\,ds = \int_0^6 (s^5 - 10s^4 + 35s^3 - 50s^2 + 24s)\,ds$$

$$= \frac{6^6}{6} - \frac{10.6^5}{5} + \frac{35.6^4}{4} - \frac{50.6^3}{3} + \frac{24.6^2}{2} = 396$$

$$\int_0^6 s(s-1)(s-2)(s-3)(s-4)(s-5)\,ds$$

$$= \int_0^6 (s^6 - 15s^5 + 85s^4 - 225s^3 + 274s^2 - 120s)\,ds$$

$$= \frac{6^7}{7} - \frac{15.6^6}{6} + \frac{85.6^5}{5} - \frac{225.6^4}{4} + \frac{274.6^3}{3} - \frac{120.6^2}{2} = \frac{1476}{7}$$

Thus, using the values of these individual integrals, we obtain the value of

the integral $\int_{x_0}^{x_6} f_0(x)\,dx = h\int_0^6 f_0(x)\,ds$ as follows:

$$\begin{aligned}
&\int_{x_0}^{x_6} f_0(x)\,dx \\
&= h\Big[6y_0+18\boldsymbol{\Delta}y_0+54\tfrac{\boldsymbol{\Delta}^2y_0}{2!}+144\tfrac{\boldsymbol{\Delta}^3y_0}{3!}+\tfrac{1476}{5}\tfrac{\boldsymbol{\Delta}^4y_0}{4!}+396\tfrac{\boldsymbol{\Delta}^5y_0}{5!}+\tfrac{1476}{7}\tfrac{\boldsymbol{\Delta}^6y_0}{6!}\Big] \\
&= h\big[6y_0+18\boldsymbol{\Delta}y_0+27\boldsymbol{\Delta}^2y_0+24\boldsymbol{\Delta}^3y_0+\tfrac{123}{10}\boldsymbol{\Delta}^4y_0+\tfrac{33}{10}\boldsymbol{\Delta}^5y_0+\tfrac{41}{140}\boldsymbol{\Delta}^6y_0\big]
\end{aligned} \tag{12.4}$$

$$\begin{aligned}
&= h[6y_0+18(y_1-y_0)+27(y_2-2y_1+y_0)+24(y_3-3y_2+3y_1-y_0) \\
&\quad +\tfrac{123}{10}(y_4-4y_3+6y_2-4y_1+y_0)+\tfrac{33}{10}(y_5-5y_4+10y_3-10y_2+5y_1-y_0) \\
&\quad +\tfrac{41}{140}(y_6-6y_5+15y_4-20y_3+15y_2-6y_1+y_0)\big] \\
&= h\big[\big(6-18+27-24+\tfrac{123}{10}-\tfrac{33}{10}+\tfrac{41}{140}\big)y_0 \\
&\quad +\big(18-54+72-\tfrac{492}{10}+\tfrac{165}{10}-\tfrac{246}{140}\big)y_1+\big(27-72+\tfrac{738}{10}-33+\tfrac{615}{140}\big)y_2 \\
&\quad +\big(24-\tfrac{492}{10}+33-\tfrac{820}{140}\big)y_3+\big(\tfrac{123}{10}-\tfrac{165}{10}+\tfrac{615}{140}\big)y_4+\big(\tfrac{33}{10}-\tfrac{246}{140}\big)y_5 \\
&\quad +\tfrac{41}{140}y_6\big]
\end{aligned} \tag{12.5}$$

$$= h\left[\frac{41}{140}y_0+\frac{216}{140}y_1+\frac{27}{140}y_2+\frac{272}{140}y_3+\frac{27}{140}y_4+\frac{216}{140}y_5+\frac{41}{140}y_6\right]$$

$$= \frac{h}{140}\left(41y_0+216y_1+27y_2+272y_3+27y_4+216y_5+41y_6\right)$$

It has to be noted that, these values of the coefficients of y_i, $i = 0, 1, \ldots, 6$ existing in the above Weddle's integral (12.5) are well stated in the page 116 of [40], but **without any proof** though.

Exactly in the same way as above, we obtain the following

$$\int_{x_6}^{x_{12}} f_0(x)\,dx = \frac{h}{140}\left(41y_6 + 216y_7 + 27y_8 + 272y_9 + 27y_{10} + 216y_{11} + 41y_{12}\right)$$

Hence, taking n to be a multiple of 6, we arrive at

$$\begin{aligned}
&\int_{x_0}^{x_n} f_0(x)\,dx \\
&= \int_{x_0}^{x_6} f_0(x)\,dx + \int_{x_6}^{x_{12}} f_0(x)\,dx + \ldots + \int_{x_{n-6}}^{x_n} f_0(x)\,dx \\
&= \frac{h}{140}\,[41y_0 + 216(y_1 + y_7 + y_{13} + \ldots + y_{n-5}) \\
&\qquad + 27(y_2 + y_8 + y_{14} + \ldots + y_{n-4}) + 272(y_3 + y_9 + y_{15} + \ldots + y_{n-3}) \\
&\qquad + 27(y_4 + y_{10} + y_{16} + \ldots + y_{n-2}) + 216(y_5 + y_{11} + y_{17} + \ldots + y_{n-1}) \\
&\qquad + 82(y_6 + y_{12} + y_{18} + \ldots + y_{n-6}) + 41y_n] \\
&= \frac{h}{140}\left[41y_0 + 216\sum_{i=0}^{\frac{n-6}{6}} y_{6i+1} + 27\sum_{i=0}^{\frac{n-6}{6}} y_{6i+2} + 272\sum_{i=0}^{\frac{n-6}{6}} y_{6i+3} + 27\sum_{i=0}^{\frac{n-6}{6}} y_{6i+4}\right. \\
&\qquad \left. +216\sum_{i=0}^{\frac{n-6}{6}} y_{6i+5} + 82\sum_{i=1}^{\frac{n-6}{6}} y_{6i} + 41y_n\right] \\
&= \frac{h}{140}\left[41f(x_0) + 216\sum_{i=0}^{\frac{n-6}{6}} f(x_{6i+1}) + 27\sum_{i=0}^{\frac{n-6}{6}} f(x_{6i+2}) + 272\sum_{i=0}^{\frac{n-6}{6}} f(x_{6i+3})\right. \\
&\qquad \left. +27\sum_{i=0}^{\frac{n-6}{6}} f(x_{6i+4}) + 216\sum_{i=0}^{\frac{n-6}{6}} f(x_{6i+5}) + +82\sum_{i=1}^{\frac{n-6}{6}} f(x_{6i}) + 41f(x_n)\right]
\end{aligned} \tag{12.6}$$

which is the **Weddle's rule of numerical integration**.

12.1.4 Estimation of error in the Weddle's integral

The exact value of the integral $\int_{x_0}^{x_6} f(x)\,dx$ is given as:

$$\int_{x_0}^{x_6} f(x)\,dx = \Phi(x_6) - \Phi(x_0) = \Phi(x_0+6h) - \Phi(x_0), \text{ where } \frac{d}{dx}\Phi(x) = f(x)$$

$$= 6hf(x_0) + \frac{(6h)^2}{2!}f^{'}(x_0) + \frac{(6h)^3}{3!}f^{''}(x_0) + \ldots + \frac{(6h)^9}{9!}f^{(8)}(x_0) + O(h^{10})$$

(applying the Taylor's theorem for $f(x)$, assuming that $f(x)$ at the point x_0 has derivatives at least up to the eighth order)

$$=6hy_0 + 18h^2y_0^{'} + 36h^3y_0^{''} + 54h^4y_0^{'''} + \frac{648}{10}h^5y_0^{(4)} + \frac{648}{10}h^6y_0^{(5)} + \frac{3888}{70}h^7y_0^{(6)} + \frac{2916}{70}h^8y_0^{(7)} + \frac{1944}{70}h^9y_0^{(8)} + O(h^{10}) \tag{12.7}$$

Again, the right hand side of the Weddle's formula (12.4) for the integral $\int_{x_0}^{x_6} f(x)\,dx \approx \int_{x_0}^{x_6} f_0(x)\,dx = W(x_0, x_6)$ reads:

$$W(x_0,x_6) = \frac{h}{140}\left[41y_0 + 216y_1 + 27y_2 + 272y_3 + 27y_4 + 216y_5 + 41y_6\right]$$

$$= \frac{h}{140}\Bigg[41y_0 + 216\left(y_0 + hy_0^{'} + \frac{h^2}{2!}y_0^{''} + \frac{h^3}{3!}y_0^{'''} + \ldots + \frac{h^8}{8!}y_0^{(8)}\right) + 27\left(y_0 + (2h)y_0^{'} + \frac{(2h)^2}{2!}y_0^{''} + \frac{(2h)^3}{3!}y_0^{'''} + \ldots + \frac{(2h)^8}{8!}y_0^{(8)}\right) + 272\left(y_0 + (3h)y_0^{'} + \frac{(3h)^2}{2!}y_0^{''} + \frac{(3h)^3}{3!}y_0^{'''} + \ldots + \frac{(3h)^8}{8!}y_0^{(8)}\right) + 27\left(y_0 + (4h)y_0^{'} + \frac{(4h)^2}{2!}y_0^{''} + \frac{(4h)^3}{3!}y_0^{'''} + \ldots + \frac{(4h)^8}{8!}y_0^{(8)}\right) + 216\left(y_0 + (5h)y_0^{'} + \frac{(5h)^2}{2!}y_0^{''} + \frac{(5h)^3}{3!}y_0^{'''} + \ldots + \frac{(5h)^8}{8!}y_0^{(8)}\right) + 41\left(y_0 + (6h)y_0^{'} + \frac{(6h)^2}{2!}y_0^{''} + \frac{(6h)^3}{3!}y_0^{'''} + \ldots + \frac{(6h)^8}{8!}y_0^{(8)}\right) + O(h^9)$$

(In each of the above steps, y_i $(i = 1, 2, \ldots, 6)$ has been expanded in form of the following finite Taylor's series:
$f(x_i) = f(x_0+ih) = y_i = y_0+(ih)y_0^{'}+\frac{(ih)^2}{2!}y_0^{''}+\frac{(ih)^3}{3!}y_0^{'''}+\ldots+\frac{(ih)^8}{8!}y_0^{(8)}+O(h^9)$,
$O(h^9)$ being the remainder after the ninth term). Hence, we get

$$\begin{aligned}
W(x_0, x_6) = \frac{h}{140}\,[&\{41 + 216 + 27 + 272 + 27 + 216 + 41\}y_0 \\
&+ \{216 + (27)2 + (272)3 + (27)4 + (216)5 + (41)6\}hy_0^{'} \\
&+ \{216 + (27)2^2 + (272)3^2 + (27)4^2 + (216)5^2 + (41)6^2\}\frac{h^2}{2}y_0^{''} \\
&+ \{216 + (27)2^3 + (272)3^3 + (27)4^3 + (216)5^3 + (41)6^3\}\frac{h^3}{6}y_0^{'''} \\
&+ \{216 + (27)2^4 + (272)3^4 + (27)4^4 + (216)5^4 + (41)6^4\}\frac{h^4}{24}y_0^{(4)} \\
&+ \{216 + (27)2^5 + (272)3^5 + (27)4^5 + (216)5^5 + (41)6^5\}\frac{h^5}{120}y_0^{(5)} \\
&+ \{216 + (27)2^6 + (272)3^6 + (27)4^6 + (216)5^6 + (41)6^6\}\frac{h^6}{720}y_0^{(6)} \\
&+ \{216 + (27)2^7 + (272)3^7 + (27)4^7 + (216)5^7 + (41)6^7\}\frac{h^7}{5040}y_0^{(7)} \\
&+ \{216 + (27)2^8 + (272)3^8 + (27)4^8 + (216)5^8 + (41)6^8\}\frac{h^8}{40320}y_0^{(8)} \\
&+ O(h^9)\big] \\
= h\Bigg[&6y_0 + 18hy_0^{'} + 36h^2y_0^{''} + 54h^3y_0^{'''} + \frac{648}{10}h^4y_0^{(4)} + \frac{648}{10}h^5y_0^{(5)} \\
&+ \frac{3888}{70}h^6y_0^{(6)} + \frac{2916}{70}h^7y_0^{(7)} + \frac{3888.9}{140}h^8y_0^{(8)}\Bigg] + O(h^{10}) \\
= &\,6hy_0 + 18h^2y_0^{'} + 36h^3y_0^{''} + 54h^4y_0^{'''} + \frac{648}{10}h^5y_0^{(4)} + \frac{648}{10}h^6y_0^{(5)} \\
&+ \frac{3888}{70}h^7y_0^{(6)} + \frac{2916}{70}h^8y_0^{(7)} + \frac{3888.9}{140}h^9y_0^{(8)} + O(h^{10})
\end{aligned} \tag{12.8}$$

Thus, using (12.7) and (12.8), we get the following error ϵ_{Weddle} as the differ-

ence between Weddle's formula $W(x_0, x_6)$ and the true value of the integral $\int_{x_0}^{x_6} f(x)\,dx$ as follows:

$$\begin{aligned}\epsilon_{Weddle} = W(x_0, x_6) - \int_{x_0}^{x_6} f(x)\,dx &= \left(\frac{3888.9}{140} - \frac{1944}{70}\right) h^9 y_0^{(8)} + O(h^{10}) \\ &= \frac{9}{1400} h^9 y_0^{(8)} + O(h^{10}) = \frac{9}{1400} h^9 f^{(8)}(x_0) + O(h^{10})\end{aligned} \tag{12.9}$$

This very expression ϵ_{Weddle} giving the estimation of error is well stated in the page 116 of [40], but **without any proof** though.

Hence, an allowable error of $O(h^9)$ makes Weddle's rule of numerical integration to be extremely useful and reliable. Of course, it is extremely important that h must be chosen to be less than 1 for each step of calculation of an integral, i.e $h < 1$, otherwise, the entire process of numerical integration would be useless.

This process of numerical integration with subject to Weddle's rule has been used to program the solution of a nonlinear system of two equations involving Riemann integrations, the descriptions of this system is given in the subsequent sections. Basically, this system involves the first two moments of the continuous random variable X having the range of variability $[0, 1]$.

12.2 Numerical solution of an equation

The equations or a system of simultaneous equations, to which we shall have to do with, do not have more than one solution each time. Therefore, our discussions will be confined to those equations, which do have only single solutions. Moreover, the functions involved in our cases are continuous, or thy have at most a finite number of removable discontinuities. A function $f(x)$ is removable discontinuous at $x = a$ means

$$\lim_{x \to a+} f(x) = \lim_{x \to a-} f(x) = l$$

For our convenience, we shall take each such l as the value of $f(x)$ at $x = a$, so as to remove all the discontinuities.

The two important relevant numerical methods for solving equations are the iterative procedure and the Newton Raphson procedure. In course of our discussions we shall see, how the iterative procedure is important for the derivation of convergence conditions for the Newton Raphson procedure.

12.2.1 Iterative procedure

Let ξ be the true value of the root of the following equation

$$f(x) = 0 \tag{12.10}$$

which can be rewritten in the following form

$$x = \phi(x) \tag{12.11}$$

such that the function $\phi(x)$ is derivable within a given neighborhood I of ξ. The description of I will be given later. Then, if $x_0 \in I$ be the first approximated value of ξ, then x_1 will be the second approximated value of ξ given by

$$x_1 = \phi(x_0) \tag{12.12}$$

Without any loss of generality, we can assume that $x_0 < \xi$. Accordingly, since $\xi = \phi(\xi)$, by the Lagrange's mean value theorem of differential calculus, there exists at least one ξ_0 such that

$$\xi - x_1 = \phi(\xi) - \phi(x_0) = (\xi - x_0)\phi^{'}(\xi_0), \quad \xi_0 \in (x_0, \xi) \tag{12.13}$$

Exactly in the same way, if x_2 be the third approximated value of ξ, then

$$\xi - x_2 = \phi(\xi) - \phi(x_1) = (\xi - x_1)\phi^{'}(\xi_1), \quad \xi_1 \in (x_1, \xi) \tag{12.14}$$

$\vdots$

Proceeding exactly in this way, if x_{n+1} be the $(n+2)^{th}$ approximated value of ξ ($n \in \mathbb{N}_0$), then

$$\begin{aligned} \xi - x_{n+1} &= \phi(\xi) - \phi(x_n) = (\xi - x_n)\phi^{'}(\xi_n), \quad \xi_n \in (x_n, \xi) && (12.15)\\ &= (\xi - x_{n-1})\phi^{'}(\xi_{n-1})\phi^{'}(\xi_n) \\ &\vdots \\ &= (\xi - x_0)\phi^{'}(\xi_0)\phi^{'}(\xi_1)\phi^{'}(\xi_2)\ldots\phi^{'}(\xi_n) && (12.16) \end{aligned}$$

It has to be noted, that in each step (ie $n \in \mathbb{N}_0$), x_{n+1} was computed with the help of x_n, by the following relation:

$$x_{n+1} = \phi(x_n) \tag{12.17}$$

Now, by taking

- $\ell = \sup_{n\in\mathbb{N}} \left|\phi^{'}(\xi_n)\right| < 1$
- $\epsilon_{n+1} = |\xi - x_{n+1}|$ = error in the $(n+2)^{th}$ approximation of the root, which means that by using (12.15) we conclude

$$\epsilon_{n+1} = |\xi - x_{n+1}| = |\xi - x_n| \left|\phi^{'}(\xi_n)\right| = \epsilon_n \left|\phi^{'}(\xi_n)\right| \tag{12.18}$$

we get the following relations

- $$|\xi - x_{n+1}| = |\xi - x_0| \left|\phi^{'}(\xi_0)\phi^{'}(\xi_1)\ldots\phi^{'}(\xi_n)\right| < |\xi - x_0|\,\ell^{n+1} \tag{12.19}$$
- $$\epsilon_{n+1} \text{ is linearly related to } \epsilon_n \tag{12.20}$$

Therefore, the inequality (12.19) clearly shows, that the sequence defined by $\{x_n\}_{n\in\mathbb{N}_0}$ converges and converges to ξ, provided $\ell < 1$. In other words, $x_n \to \xi$ as $n \to \infty$, provided $\ell < 1$.

12.2.2 Condition for the convergence of the iterative procedure

By careful consideration of the following relations

$$\begin{aligned} x_n &\rightarrow \xi \text{ as } n \rightarrow \infty \\ \xi - x_{n+1} &= (\xi - x_n)\phi^{'}(\xi_n) \end{aligned}$$

we can only conclude, that $\left|\phi^{'}(\xi_n)\right| < 1$ for every $n \in \mathbb{N}_0$, since it is evident that

- x_{n+1} is closer to ξ than that of x_n

Therefore, we can condition the convergence of the sequence $\{x_n\}_{n\in\mathbb{N}_0}$ by asserting $\ell < 1$.

Therefore, since x_0 is undoubtedly the farthest point from ξ than that of the other iterated points $x_1, x_2, \ldots, x_n, \ldots$, we conclude:

The sequence $\{x_n\}_{n\in\mathbb{N}_0}$ can be made to converge and converge to ξ, if we set $\left|\phi^{'}(x_0)\right| < 1$.

Whence, the condition for convergence for the iterative procedure is given as

$$\left|\phi^{'}(x_0)\right| < 1 \tag{12.21}$$

and the required neighborhood I of ξ be defined as

$$I = \left\{x : \quad \left|\phi^{'}(x)\right| < 1\right\} \tag{12.22}$$

which is at the same time, the interval of convergence for the iterative procedure. This is referred to the page 212 of [36].

12.2.3 Rate of convergence of the iterative procedure

The relation (12.15) shows the linearity of the relationship between errors in two successive approximations of the root ξ and (12.20) confirms the same. This shows that the iterative procedure has a convergence of first order or a linear convergence.

12.2.4 Newton Raphson procedure

If h be the error in the root ξ of the equation (12.10), such that $x + h = \xi$, then we get

$$f(x + h) = 0 \tag{12.23}$$

Now, the Newton's assumptions say, that

- There exists a $\theta \in (0, 1)$, such that $f(x+h)$ can be expanded by Taylor's theorem about x as

$$f(x + h) = f(x) + hf^{'}(x) + \frac{h^2}{2}f^{''}(x + \theta h) \tag{12.24}$$

- The first approximated value x_0 of ξ, such that $x_0 + h = \xi$, is chosen sufficiently close to ξ in the sense, that the term $\frac{h^2}{2}f^{''}(x_0 + \theta h)$ is small enough to make sure, that the Newton Raphson procedure ensures the success of finding the solution of (12.10). This smallness is explainable by the convergence condition derived in next subsection. This brings us to

$$0 = f(x_0 + h) \approx f(x_0) + hf^{'}(x_0) \Rightarrow h \approx -\frac{f(x_0)}{f^{'}(x_0)} \tag{12.25}$$

With the help of the approximated value of h given in (12.25), the second approximated value x_1 of ξ is given as

$$x_1 = x_0 - \frac{f(x_0)}{f^{'}(x_0)} \tag{12.26}$$

Exactly in the same way, the third approximated value x_2 of ξ is given as

$$x_2 = x_1 - \frac{f(x_1)}{f^{'}(x_1)} \tag{12.27}$$

$\vdots$

Proceeding exactly in this way, if x_{n+1} be the $(n + 2)^{th}$ approximated value of ξ ($n \in \mathbb{N}_0$), then

$$x_{n+1} = x_n - \frac{f(x_n)}{f^{'}(x_n)} \tag{12.28}$$

Thus, the procedure of determination of such x_{n+1}, which is a sufficiently good approximation of the true solution of (12.10) is known as the Newton Raphson procedure

(12.29)

Now, the question arises, whether the usage of the Newton Raphson procedure (12.29) really leads us to our desired result $x_n \to \xi$ as $n \to \infty$ or not.

For this, we shall have to derive the condition of the convergence of the Newton Raphson procedure.

12.2.5 Condition for the convergence of the Newton Raphson procedure

By comparing the relations (12.17) and (12.29), viz

$$\begin{cases} x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \\ x_{n+1} = \phi(x_n) \end{cases}$$

we get $\phi(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}$ which leads us to

$$\phi(x) = x - \frac{f(x)}{f'(x)} \tag{12.30}$$

$$\Rightarrow \quad \phi'(x) = \frac{f(x)f''(x)}{\left(f'(x)\right)^2} \tag{12.31}$$

With the help of (12.21), the last relation leads us to conclude the following:

The condition for convergence for the Newton Raphson procedure is given as

$$\left|\phi'(x_0)\right| = \frac{\left|f(x_0)f''(x_0)\right|}{\left(f'(x_0)\right)^2} < 1 \tag{12.32}$$

The fulfillment of the above condition ensures the necessary smallness of the said $\frac{h^2}{2}f''(x_0 + \theta h)$ in (12.24). In other words, the degree of smallness of $\frac{h^2}{2}f''(x_0 + \theta h)$ determined by the fulfillment of condition (12.32) enabled us to neglect this term (i.e. $\frac{h^2}{2}f''(x_0 + \theta h)$) before we proceeded from (12.24).

12.2.6 Rate of convergence of the Newton Raphson procedure

By putting $x = x_n$ and $h = \xi - x_n$ in (12.23) and (12.24), we get

$$\begin{aligned} 0 &= f(x_n+\xi-x_n)=f(x_n)+(\xi-x_n)f^{'}(x_n)+\tfrac{1}{2}(\xi-x_n)^2 f^{''}(x_n+\theta(\xi-x_n)) \\ &\Rightarrow -\frac{f(x_n)}{f^{'}(x_n)} = (\xi - x_n) + \frac{1}{2}(\xi - x_n)^2 \frac{f^{''}(x_n + \theta(\xi - x_n))}{f^{'}(x_n)} \\ &\Rightarrow x_n - \frac{f(x_n)}{f^{'}(x_n)} = \xi + \frac{1}{2}(\xi - x_n)^2 \frac{f^{''}(x_n + \theta(\xi - x_n))}{f^{'}(x_n)} \end{aligned} \tag{12.33}$$

Then, by putting $x_{n+1} = x_n - \frac{f(x_n)}{f^{'}(x_n)}$ in the above relation, we get

$$\begin{aligned} x_{n+1} &= \xi + \frac{1}{2}(\xi - x_n)^2 \frac{f^{''}(x_n + \theta(\xi - x_n))}{f^{'}(x_n)} \\ \Rightarrow \quad x_{n+1} - \xi &= \frac{1}{2}(\xi - x_n)^2 \frac{f^{''}(x_n + \theta(\xi - x_n))}{f^{'}(x_n)} \end{aligned} \tag{12.34}$$

Again, by using the definition of the error in the $(n+2)^{th}$ approximation of the root ξ given by $\epsilon_{n+1} = |x_{n+1} - \xi|$, we get the above relation as

$$\epsilon_{n+1} = \frac{1}{2}\epsilon_n^2 \left| \frac{f^{''}(x_n + \theta(\xi - x_n))}{f^{'}(x_n)} \right| \tag{12.35}$$

This shows, that the Newton Raphson procedure has a convergence of second order or quadratic convergence. This means, as ϵ_{n+1} is connected with the square of ϵ_n, the smallness determined by ϵ_n in the n^{th} has been squared to give a squared smallness in the $n + 1^{\text{th}}$ step.

It is clear, that the convergence speed of the Newton Raphson procedure is undoubtedly higher than the same of the iterative procedure. This justifies the usage of the Newton Raphson procedure for solving an equation with subject to the fulfillment of the condition (12.32).

12.2.7 Complete solution of the equation $f(x) = 0$

Before we actually use the Newton Raphson procedure to solve the equation (12.10), we need to find an appropriate x_0 which satisfies the condition

(12.32). In other words, in order to find a reasonably accurate value of ξ, we must find the first approximated value x_0 of ξ satisfying the (12.32). For to do that, we shall have to go through the following algorithmic steps:

1. Location of ξ needs to be found at first and for that,

   - let $z_1 = z_2 = 0$ and calculate the value of $f(0)$
   - while $(f(z_1)f(z_2) \geq 0)$
     {
     $z_1 = z_1 - 1$; $z_2 = z_2 + 1$
     }

2. A suitable x_0 satisfying the Newton's convergence condition must be found by the method of bisection in the following manner,

   - let $z_0 = \frac{z_1+z_2}{2}$ and $x_0 = z_1$;
     compute $\phi^{'}(x_0)$ as stated in (12.32) and examine it's fulfillment.
   - while (x_0 does not fulfill the convergence condition (12.32))
     {
     if $f(z_1)f(z_0) <= 0$ then { $z_2 = z_0$; $z_0 = \frac{z_1+z_2}{2}$ };
     if $f(z_2)f(z_0) <= 0$ then { $z_1 = z_0$; $z_0 = \frac{z_1+z_2}{2}$ };
     $x_0 = z_1$;
     compute $\phi^{'}(x_0)$ as stated in (12.32) and examine it's fulfillment.
     }
     x_0 is the final first approximation of ξ, fulfilling the convergence conditions.

3. For any arbitrarily chosen $\epsilon > 0$ and by using this x_0, the value of ξ to the desired level of accuracy stated by $|x_n - \xi| < \epsilon$ is reached by the successive usage of the Newton Raphson rule (12.29)

(12.36)

Remarks:

- Referring to the first step of our above procedure for the solution, we have initialized the variables z_1 and z_2 with zero: $z_1 = z_2 = 0$. Instead of zero, if we are in a position to initialize the same with another real

number $\overline{\xi}$ such that $\left|\overline{\xi} - \xi\right| < |\xi|$, then we can undoubtedly lower the running time of the program. The only question is of the availability of such $\overline{\xi}$.

- Our programming outputs show, that if $|x_0 - \xi| < 10^{-10}$ is fulfilled, then the Newton's convergence condition for solving the equations relevant for our interests is fulfilled.

12.3 Numerical solution of a system of two equations

We have seen in the previous section, that the Newton Raphson procedure for solving an equation with a single unknown has a speedy convergence. In this section, our problem will be to solve a simultaneous system of two equations with two unknowns. For to solve this problem, the Newton Raphson procedure can be extended to two unknowns.

12.3.1 Newton Raphson procedure

Let $(x_0 + h, y_0 + k)$ be the true solution of the following system of equations

$$\begin{cases} f(x, y) = 0 \\ g(x, y) = 0 \end{cases} \tag{12.37}$$

such that (x_0, y_0) is the first approximated solution of the system (12.37) and the values h and k are the errors in the first and the second components of the true solution respectively.

Now, the Newton's assumptions say, that

- By expanding both the functions $f(x, y)$ and $g(x, y)$ by Taylor's theorem about (x_0, y_0), we get

$$\begin{cases} f(x_0 + h, y_0 + k) = f(x_0, y_0) + hf_x(x_0, y_0) + kf_y(x_0, y_0) + T_f(h, k) = 0 \\ g(x_0 + h, y_0 + k) = g(x_0, y_0) + hg_x(x_0, y_0) + kg_y(x_0, y_0) + T_g(h, k) = 0 \end{cases} \tag{12.38}$$

  where $T_f(h, k)$ and $T_g(h, k)$ are the expressions containing the terms in h or k (or both) of powers higher than or equal to two and f_x, g_x, f_y and g_y are partial derivatives having their usual meanings.

- The first approximated solution (x_0, y_0) is chosen sufficiently close to (x_0+h, y_0+k) in the sense, that the terms $T_f(h, k)$ and $T_g(h, k)$ are small enough to ensure, that the Newton Raphson procedure ensures the success of finding the solution of (12.37). This smallness is explainable by convergence conditions stated in the next subsection. This brings us to

$$\begin{cases} hf_x(x_0, y_0) + kf_y(x_0, y_0) \approx -f(x_0, y_0) \\ hg_x(x_0, y_0) + kg_y(x_0, y_0) \approx -g(x_0, y_0) \end{cases}$$

Therefore, the approximated values of h and k can be found out by solving the following system of linear equations

$$\begin{cases} hf_x(x_0, y_0) + kf_y(x_0, y_0) = -f(x_0, y_0) \\ hg_x(x_0, y_0) + kg_y(x_0, y_0) = -g(x_0, y_0) \end{cases} \tag{12.39}$$

With the help of the approximated values of h and k, namely h_0 and k_0 as a result of the solution of the system (12.39), the second approximated solution (x_1, y_1) of the system (12.37) is given as

$$\begin{cases} x_1 = x_0 + h_0 \\ y_1 = y_0 + k_0 \end{cases} \tag{12.40}$$

Exactly in the same way, the third approximated solution (x_2, y_2) of the system (12.37) is given as

$$\begin{cases} x_2 = x_1 + h_1 \\ y_2 = y_1 + k_1 \end{cases} \tag{12.41}$$

where h_1 and k_1 are the approximated values of h and k respectively as a result of the solution of the system

$$\begin{cases} hf_x(x_1, y_1) + kf_y(x_1, y_1) = -f(x_1, y_1) \\ hg_x(x_1, y_1) + kg_y(x_1, y_1) = -g(x_1, y_1) \end{cases} \tag{12.42}$$

$$\vdots$$

Proceeding exactly in this way, the $(n+2)^{th}$ approximated solution (x_{n+1}, y_{n+1}) of the system (12.37) is given as

$$\begin{cases} x_{n+1} = x_n + h_n \\ y_{n+1} = y_n + k_n \end{cases} \tag{12.43}$$

where h_n and k_n are the approximated values of h and k respectively as a result of the solution of the system

$$\begin{cases} hf_x(x_n, y_n) + kf_y(x_n, y_n) = -f(x_n, y_n) \\ hg_x(x_n, y_n) + kg_y(x_n, y_n) = -g(x_n, y_n) \end{cases} \tag{12.44}$$

Expectedly, (x_{n+1}, y_{n+1}) is supposed to go arbitrarily near to the true solution of (12.37) with the increase in $n \in \mathbb{N}$.

Thus, the procedure of determination of such (x_{n+1}, y_{n+1}), which is a sufficiently good approximation of the true solution of (12.37) is known as the Newton Raphson procedure.

Hence the **Newton Raphson procedure for two unknowns** can be formally given by rewriting the system of equations (12.44) in the matrix notation as follows:

$$\begin{pmatrix} h \\ k \end{pmatrix} = - \begin{pmatrix} f_x(x_n, y_n) & f_y(x_n, y_n) \\ g_x(x_n, y_n) & g_y(x_n, y_n) \end{pmatrix}^{-1} \begin{pmatrix} f \\ g \end{pmatrix} \tag{12.45}$$

Moreover, situations may arise, when the Newton Raphson procedure described by (12.45) may need a suitable refinement to ensure the security of the direction of convergence. For this, a damping factor t, $0 < t \leq 1$ is introduced to (12.45), so that

$$\begin{pmatrix} h \\ k \end{pmatrix} = -t \begin{pmatrix} f_x(x_n, y_n) & f_y(x_n, y_n) \\ g_x(x_n, y_n) & g_y(x_n, y_n) \end{pmatrix}^{-1} \begin{pmatrix} f \\ g \end{pmatrix} \tag{12.46}$$

In that case, the Newton Raphson procedure described by (12.46) is known as the **damped Newton Raphson procedure for two unknowns**.

Now, the question arises, what could be the conditions, under which the described Newton Raphson procedure (12.45) or the described damped Newton Raphson procedure (12.46) with subject to suitably chosen t really leads us to the solution of the system (12.37).

For this, we shall have to take the necessary conditions of the convergence of the Newton Raphson procedure into our account.

With subject to the consideration of the systems (11.37) and (11.58), we shall refer to the **theorem 5.3.2**, which is well stated and proved in the pages 249 - 253 of [40]. This theorem 5.3.2 gives the **necessary** and **sufficient** condition for the convergence of the Newton Raphson procedure (i. e. the general **Newton Raphson's convergence criterion**) in case of n ($n \in \mathbb{N}$) unknowns. For the sake of the very best degree of clarity, **we shall state and prove this theorem 5.3.2 formally and rigorously**, before we go ahead to discuss the role of Newton Raphson procedure in our systems of simultaneous equations (11.37) and (11.58).

This rigorous proof of the necessary and the sufficient condition of convergence necessitates the proof of a lemma at first.

12.3.2 The lemma for the convergence of the Newton Raphson procedure

The statement of the lemma for the Newton Raphson's convergence criterion: If the derivative of the function $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})$, namely $\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})$, exists for every $\overrightarrow{\mathbf{x}} \in \mathbf{D}_0$, such that $\mathbf{D}_0$ is a convex set with $\mathbf{D}_0 \subseteq \mathbb{R}^n$ and there exists a real valued constant $\mathbf{c}$ so that

$$\left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}} + \overrightarrow{\mathbf{h_x}}) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}) \right\| \leq \mathbf{c}\|\overrightarrow{\mathbf{h_x}}\| \quad \textit{for every} \quad \overrightarrow{\mathbf{x}}, \overrightarrow{\mathbf{y}} \in \mathbf{D}_0, \\ \textit{such that } \overrightarrow{\mathbf{y}} = \overrightarrow{\mathbf{x}} + \overrightarrow{\mathbf{h_x}} \tag{12.54}$$

holds, then an upper bound of the expression

$$\left\| \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}} + \overrightarrow{\mathbf{h_x}}) - \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})\, \overrightarrow{\mathbf{h_x}} \right\|$$

can be suitably given by the following inequality:

$$\left\| \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}} + \overrightarrow{\mathbf{h_x}}) - \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})\, \overrightarrow{\mathbf{h_x}} \right\| \leq \frac{\mathbf{c}}{2}\|\overrightarrow{\mathbf{h_x}}\|^2 \tag{12.47}$$

Proof of the lemma: By defining a differentiable function $\overrightarrow{\phi} : [0,1] \to \mathbb{R}^n$ within it's domain of definition $[0,1]$ by $\overrightarrow{\phi}(t) = \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}} + t\, \overrightarrow{\mathbf{h_x}})$, we get by using the chain rule of differentiation the derivative of $\overrightarrow{\phi}(t)$ as:

$$\overrightarrow{\phi}'(t) = \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}} + t\, \overrightarrow{\mathbf{h_x}})\, \overrightarrow{\mathbf{h_x}} \tag{12.48}$$

(in plain words, we have differentiated $\overrightarrow{\phi}(t)$ with respect to t, $0 \leq t \leq 1$ by keeping $\overrightarrow{\mathbf{x}}$, $\overrightarrow{\mathbf{x}} + \overrightarrow{\mathbf{h_x}} \in \mathbf{D}_0$ in mind.)

In that case, by using the above definition of $\overrightarrow{\phi}'(t)$, namely by (12.48), we get

$$\begin{aligned} \overrightarrow{\phi}'(t) - \overrightarrow{\phi}'(0) &= \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}} + t\overrightarrow{\mathbf{h_x}})\, \overrightarrow{\mathbf{h_x}} - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})\, \overrightarrow{\mathbf{h_x}} \\ &= \left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}} + t\overrightarrow{\mathbf{h_x}}) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}) \right) \overrightarrow{\mathbf{h_x}} \end{aligned} \tag{12.49}$$

Here, by using the above expressions of both $\overrightarrow{\phi}(t)$ and $\overrightarrow{\phi}'(t)$, we arrive at

$$\begin{aligned}\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}+\overrightarrow{\mathbf{h}_{\mathbf{x}}})-\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})-\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})\,\overrightarrow{\mathbf{h}_{\mathbf{x}}} &= \overrightarrow{\phi}(1)-\overrightarrow{\phi}(0)-\overrightarrow{\phi}'(0)\\ &= \int_0^1\left(\overrightarrow{\phi}'(t)-\overrightarrow{\phi}'(0)\right)dt \qquad (12.50)\end{aligned}$$

and consequently

$$\begin{aligned}&\left\|\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}+\overrightarrow{\mathbf{h}_{\mathbf{x}}})-\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})-\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})\,\overrightarrow{\mathbf{h}_{\mathbf{x}}}\right\|\\ &=\left\|\int_0^1\left(\overrightarrow{\phi}'(t)-\overrightarrow{\phi}'(0)\right)dt\right\|\\ &\le\int_0^1\left\|\overrightarrow{\phi}'(t)-\overrightarrow{\phi}'(0)\right\|dt\\ &=\int_0^1\left\|\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}+t\overrightarrow{\mathbf{h}_{\mathbf{x}}})-\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})\right)\,\overrightarrow{\mathbf{h}_{\mathbf{x}}}\right\|dt\\ &(\text{ by using (12.49) })\\ &\le\int_0^1\left\|\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}+t\overrightarrow{\mathbf{h}_{\mathbf{x}}})-\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})\right)\right\|\left\|\overrightarrow{\mathbf{h}_{\mathbf{x}}}\right\|dt\end{aligned}$$

$$\le\int_0^1\mathbf{c}\left\|t\overrightarrow{\mathbf{h}_{\mathbf{x}}}\right\|\left\|\overrightarrow{\mathbf{h}_{\mathbf{x}}}\right\|dt$$

(by using (12.54), where $\overrightarrow{\mathbf{x}}+\overrightarrow{\mathbf{h}_{\mathbf{x}}}\in\mathbf{D}_0\Rightarrow\overrightarrow{\mathbf{x}}+t\overrightarrow{\mathbf{h}_{\mathbf{x}}}\in\mathbf{D}_0$ for $0\le t\le 1$)

$$=\int_0^1\mathbf{c}\;t\;\|\overrightarrow{\mathbf{h}_{\mathbf{x}}}\|^2dt=\frac{\mathbf{c}}{2}\|\overrightarrow{\mathbf{h}_{\mathbf{x}}}\|^2 \qquad \square$$

and **this proves the lemma** stated by (12.47).

12.3.3 Condition for the convergence of the Newton Raphson procedure

The statement of the Newton Raphson's convergence criterion: *Let* $\mathbf{D}$ *be an open subset of* $\mathbb{R}^n$ *(*$\mathbf{D} \subseteq \mathbb{R}^n$ *with* $n \in \mathbb{N}$*) and let* $\mathbf{D}_0$ *be a convex set, such that* $\overline{\mathbf{D}}_0 \subseteq \mathbf{D}$*. Let the function* $\overrightarrow{\mathbf{f}} : \mathbf{D} \mapsto \mathbb{R}^n$ *be derivable for every* $\overrightarrow{\mathbf{x}} \in \mathbf{D}_0$ *and be continuous for every* $\overrightarrow{\mathbf{x}} \in \mathbf{D}$*.*

In that case, if for a vector $\overrightarrow{\mathbf{x}}_0 \in \mathbf{D}_0$*, there exist certain positive constants* $\mathbf{r}$*,* $\mathbf{a}$*,* $\mathbf{b}$*,* $\mathbf{c}$ *and* $\mathbf{h}$ *with the following properties:*

$$S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0) = \left\{ \overrightarrow{\mathbf{x}} \,\middle|\, \|\overrightarrow{\mathbf{x}} - \overrightarrow{\mathbf{x}}_0\| < \mathbf{r} \right\} \subseteq \mathbf{D}_0 \tag{12.51}$$

$$\mathbf{h} = \frac{\mathbf{a}\,\mathbf{b}\,\mathbf{c}}{2} < 1 \tag{12.52}$$

$$\mathbf{r} = \frac{\mathbf{a}}{1 - \mathbf{h}} \tag{12.53}$$

and if $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})$ *happens to have the following properties:*

- (a) $\left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}} + \overrightarrow{\mathbf{h}_{\mathbf{x}}}) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}) \right\| \leq \mathbf{c}\|\overrightarrow{\mathbf{h}_{\mathbf{x}}}\|$ *for every* $\overrightarrow{\mathbf{x}}, \overrightarrow{\mathbf{y}} \in \mathbf{D}_0$*, such that* $\overrightarrow{\mathbf{y}} = \overrightarrow{\mathbf{x}} + \overrightarrow{\mathbf{h}_{\mathbf{x}}}$

  (12.54)

- (b) $\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}) \right)^{-1}$ *exists and* $\left\| \left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}) \right)^{-1} \right\| \leq \mathbf{b}$ *for every* $\overrightarrow{\mathbf{x}} \in \mathbf{D}_0$

  (12.55)

- (c) $\left\| \left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0) \right)^{-1} \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_0) \right\| \leq \mathbf{a}$

  (12.56)

then the following do hold good and are required to be proved:

- (**A**) *Starting with* $\overrightarrow{\mathbf{x}}_0$, *every vector* $\overrightarrow{\mathbf{x}}_{k+1} = \overrightarrow{\mathbf{x}}_k - \left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k)\right)^{-1} \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k)$ *is well defined and* $\overrightarrow{\mathbf{x}}_k \in S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0)$ *holds for every* $k \in \mathbb{N}_0$.

(12.57)

- (**B**) $\lim\limits_{k\to\infty} \overrightarrow{\mathbf{x}}_k = \overrightarrow{\xi}$ *exists, such that* $\overrightarrow{\xi} \in \overline{S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0)}$, *for which* $\overrightarrow{\mathbf{f}}(\overrightarrow{\xi}) = \overrightarrow{0}$

(12.58)

- (**C**) *for every* $k \in \mathbb{N}_0$, $\left\| \overrightarrow{\mathbf{x}}_k - \overrightarrow{\xi} \right\| < \mathbf{a}\frac{\mathbf{h}^{2^k-1}}{1-\mathbf{h}^{2^k}}$

(12.59)

It should be well noted that, because of the very fact that $0 < \mathbf{h} < 1$, *the speed of the convergence of the Newton Raphson procedure, at the very least, is quadratic.*

The proof of the Newton's convergence criterion, namely (**A**), (**B**) and (**C**), i.e. the fulfillments of (12.57), (12.58) and (12.59) by the function $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})$ necessitate the proof of the lemma (12.47). With this, we shall go ahead to prove the **Newton Raphson's convergence criterion**.

Proof: According to (**A**), namely (12.57) with subject to $k \in \mathbb{N}_0$, we have

$$\overrightarrow{\mathbf{x}}_{k+1} = \overrightarrow{\mathbf{x}}_k - \left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k)\right)^{-1} \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k) \Leftrightarrow \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k)(\overrightarrow{\mathbf{x}}_{k+1} - \overrightarrow{\mathbf{x}}_k) = -\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k) \tag{12.60}$$

from which we easily get

$$\begin{aligned}
\left\|\overrightarrow{\mathbf{h}}_k\right\| = \left\|\overrightarrow{\mathbf{x}}_{k+1} - \overrightarrow{\mathbf{x}}_k\right\| &= \left\|\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k)\right)^{-1} \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k)\right\| \\
&\leq \left\|\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k)\right)^{-1}\right\| \left\|\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k)\right\| \\
&\leq \mathbf{b} \left\|\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k)\right\| \text{ (by using the } \textbf{property (b)}\text{, namely (12.55))} \\
&= \mathbf{b} \left\|\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k) - \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_{k-1}) + \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_{k-1})\right\| \\
&= \mathbf{b} \left\|\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k) - \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_{k-1}) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_{k-1})(\overrightarrow{\mathbf{x}}_k - \overrightarrow{\mathbf{x}}_{k-1})\right\| \\
&\quad \text{(by using the right hand side of (12.60)), but for } k \geq 1 \\
&\leq \mathbf{b}\,\frac{\mathbf{c}}{2} \left\|\overrightarrow{\mathbf{x}}_k - \overrightarrow{\mathbf{x}}_{k-1}\right\|^2 = \frac{\mathbf{b}\,\mathbf{c}}{2} \left\|\overrightarrow{\mathbf{h}}_{k-1}\right\|^2 \\
&\quad \text{(by using the } \textbf{established lemma}\text{, namely (12.47))} \\
&= \frac{\mathbf{h}}{\mathbf{a}} \left\|\overrightarrow{\mathbf{h}}_{k-1}\right\|^2 \\
&\quad \text{(by using the property (12.52))}
\end{aligned} \tag{12.61}$$

So, we have established that $\overrightarrow{\mathbf{h}}_k$, being defined by $\overrightarrow{\mathbf{h}}_k = \overrightarrow{\mathbf{x}}_{k+1} - \overrightarrow{\mathbf{x}}_k$, is related to $\overrightarrow{\mathbf{h}}_{k-1}$ recursively and therefore by using the immediately aforesaid recursive relationship, namely (12.61), we get

$$\begin{aligned}
\left\|\overrightarrow{\mathbf{h}}_k\right\| &\leq \frac{\mathbf{h}}{\mathbf{a}}\left\|\overrightarrow{\mathbf{h}}_{k-1}\right\|^2 \leq \frac{\mathbf{h}}{\mathbf{a}}\left(\frac{\mathbf{h}}{\mathbf{a}}\left\|\overrightarrow{\mathbf{h}}_{k-2}\right\|^2\right)^2 = \left(\frac{\mathbf{h}}{\mathbf{a}}\right)\left(\frac{\mathbf{h}}{\mathbf{a}}\right)^2\left\|\overrightarrow{\mathbf{h}}_{k-2}\right\|^{2^2} \\
&\leq \left(\frac{\mathbf{h}}{\mathbf{a}}\right)\left(\frac{\mathbf{h}}{\mathbf{a}}\right)^2\left(\frac{\mathbf{h}}{\mathbf{a}}\right)^{2^2}\left\|\overrightarrow{\mathbf{h}}_{k-3}\right\|^{2^3} \\
&\vdots \\
&\leq \left(\frac{\mathbf{h}}{\mathbf{a}}\right)\left(\frac{\mathbf{h}}{\mathbf{a}}\right)^2\left(\frac{\mathbf{h}}{\mathbf{a}}\right)^{2^2}\cdots\left(\frac{\mathbf{h}}{\mathbf{a}}\right)^{2^{k-1}}\left\|\overrightarrow{\mathbf{h}}_{k-k}\right\|^{2^k} \\
&= \left(\frac{\mathbf{h}}{\mathbf{a}}\right)^{1+2+2^2+2^3+\ldots+2^{k-1}}\left\|\overrightarrow{\mathbf{h}}_0\right\|^{2^k} \\
&= \left(\frac{\mathbf{h}}{\mathbf{a}}\right)^{2^k-1}\left\|\overrightarrow{\mathbf{x}}_1 - \overrightarrow{\mathbf{x}}_0\right\|^{2^k} \\
&= \left(\frac{\mathbf{h}}{\mathbf{a}}\right)^{2^k-1}\left\|-\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0)\right)^{-1}\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_0)\right\|^{2^k} \\
&\qquad (\text{ by using the left hand side of (12.60) }), \text{ for } k=0 \\
&= \left(\frac{\mathbf{h}}{\mathbf{a}}\right)^{2^k-1}\left\|\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0)\right)^{-1}\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_0)\right\|^{2^k} \\
&\leq \left(\frac{\mathbf{h}}{\mathbf{a}}\right)^{2^k-1}\mathbf{a}^{2^k} \\
&\qquad (\text{ by using the } \textbf{property (c)}, \text{ namely (12.56) }) \\
&= \mathbf{a}\,\mathbf{h}^{2^k-1}
\end{aligned} \tag{12.62}$$

which evidently proves that the **Euclidean distance** between the vectors $\overrightarrow{\mathbf{x}}_{k+1}$ and $\overrightarrow{\mathbf{x}}_k$, namely the quantity $\left\|\overrightarrow{\mathbf{h}}_k\right\|$ (i.e. the magnitude of the distance between $\overrightarrow{\mathbf{x}}_{k+1}$ and $\overrightarrow{\mathbf{x}}_k$) is **bounded above** for every $k \in \mathbb{N}_0$, because of the very fact that $0 < \mathbf{h} < 1$.

At this very point, we are in a position to prove that the point (i.e. the vector) $\overrightarrow{\mathbf{x}}_{k+1}$ belongs to the stated neighbourhood of $\overrightarrow{\mathbf{x}}_0$, namely the set $S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0)$.

Keeping this in mind, we proceed to prove this very assertion by making use of (12.62) as follows:

$$\begin{aligned}
\left\|\overrightarrow{\mathbf{x}}_{k+1}-\overrightarrow{\mathbf{x}}_0\right\| &= \left\|\overrightarrow{\mathbf{x}}_{k+1}-\overrightarrow{\mathbf{x}}_k+\overrightarrow{\mathbf{x}}_k-\overrightarrow{\mathbf{x}}_{k-1}+\ldots+\overrightarrow{\mathbf{x}}_1-\overrightarrow{\mathbf{x}}_0\right\| \\
&\leq \left\|\overrightarrow{\mathbf{x}}_{k+1}-\overrightarrow{\mathbf{x}}_k\right\|+\left\|\overrightarrow{\mathbf{x}}_k-\overrightarrow{\mathbf{x}}_{k-1}\right\|+\ldots+\left\|\overrightarrow{\mathbf{x}}_1-\overrightarrow{\mathbf{x}}_0\right\| \\
&\leq \mathbf{a}\mathbf{h}^{2^k-1}+\mathbf{a}\mathbf{h}^{2^{k-1}-1}+\ldots+\mathbf{a}\mathbf{h}^{2^2-1}+\mathbf{a}\mathbf{h}^{2^1-1}+\mathbf{a}\mathbf{h}^{2^0-1} \\
&\qquad (\text{ by using (12.62) for all the terms }) \\
&= \mathbf{a}\left(\mathbf{h}^{2^0-1}+\mathbf{h}^{2^1-1}+\mathbf{h}^{2^2-1}+\ldots+\mathbf{h}^{2^{k-1}-1}+\mathbf{h}^{2^k-1}\right) \\
&< \mathbf{a}\left(1+\mathbf{h}+\mathbf{h}^2+\mathbf{h}^3+\ldots\infty\right) \\
&\qquad (\text{ since, by the property (12.52), } 0<\mathbf{h}<1\ ) \\
&= \frac{\mathbf{a}}{1-\mathbf{h}}=\mathbf{r} \\
&\qquad (\text{ by the statement (12.53) })
\end{aligned}$$

which means nothing, but

$$\left\|\overrightarrow{\mathbf{x}}_{k+1}-\overrightarrow{\mathbf{x}}_0\right\|<\mathbf{r} \text{ for every } k\in\mathbb{N}_0 \tag{12.63}$$

and this proves our very assertion that $\overrightarrow{\mathbf{x}}_{k+1}$ belongs to the neighbourhood $S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0)$ of $\overrightarrow{\mathbf{x}}_0$. In other words,

$$\overrightarrow{\mathbf{x}}_{k+1}\in S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0) \text{ for every } k\in\mathbb{N}_0 \tag{12.64}$$

and thereby proving the statement (**A**) fully, namely (12.57).

Nextly, we show that the sequence $\{\overrightarrow{\mathbf{x}}_k\}_{k\in\mathbb{N}_0}$ is a **Cauchy sequence**, i.e. a **convergence sequence**. So, for every $n,k\in\mathbb{N}_0$, we get

$$\begin{aligned}
&\left\|\overrightarrow{\mathbf{x}}_{n+k}-\overrightarrow{\mathbf{x}}_n\right\| \\
&\quad= \left\|\overrightarrow{\mathbf{x}}_{n+k}-\overrightarrow{\mathbf{x}}_{n+k-1}+\overrightarrow{\mathbf{x}}_{n+k-1}-\overrightarrow{\mathbf{x}}_{n+k-2}+\ldots+\overrightarrow{\mathbf{x}}_{n+1}-\overrightarrow{\mathbf{x}}_n\right\| \\
&\quad\leq \left\|\overrightarrow{\mathbf{x}}_{n+k}-\overrightarrow{\mathbf{x}}_{n+k-1}\right\|+\left\|\overrightarrow{\mathbf{x}}_{n+k-1}-\overrightarrow{\mathbf{x}}_{n+k-2}\right\|+\ldots+\left\|\overrightarrow{\mathbf{x}}_{n+1}-\overrightarrow{\mathbf{x}}_n\right\| \\
&\quad= \left\|\overrightarrow{\mathbf{h}}_{n+k-1}\right\|+\left\|\overrightarrow{\mathbf{h}}_{n+k-2}\right\|+\ldots+\left\|\overrightarrow{\mathbf{h}}_n\right\| \\
&\quad= \left\|\overrightarrow{\mathbf{h}}_n\right\|+\left\|\overrightarrow{\mathbf{h}}_{n+1}\right\|+\ldots+\left\|\overrightarrow{\mathbf{h}}_{n+k-1}\right\| \\
&\quad\leq \mathbf{a}\mathbf{h}^{2^n-1}+\mathbf{a}\mathbf{h}^{2^{n+1}-1}+\mathbf{a}\mathbf{h}^{2^{n+2}-1}+\ldots+\mathbf{a}\mathbf{h}^{2^{n+k-1}-1} \\
&\qquad (\text{ by using (12.62) for all the terms }) \\
&\quad= \frac{\mathbf{a}}{\mathbf{h}}\left(\mathbf{h}^{2^n}+\mathbf{h}^{2^n\,2}+\mathbf{h}^{2^n\,4}+\mathbf{h}^{2^n\,8}+\ldots+\mathbf{h}^{2^n\,2^{k-1}}\right)
\end{aligned}$$

and consequently

$$\begin{aligned}\left\|\overrightarrow{\mathbf{x}}_{n+k}-\overrightarrow{\mathbf{x}}_{n}\right\| &\leq \frac{\mathbf{a}}{\mathbf{h}}\left(\mathbf{h}^{2^n}+\mathbf{h}^{2^n\,2}+\mathbf{h}^{2^n\,4}+\mathbf{h}^{2^n\,8}+\ldots+\mathbf{h}^{2^n\,2^{k-1}}\right)\\ &< \frac{\mathbf{a}}{\mathbf{h}}\left(\mathbf{h}^{2^n}+\mathbf{h}^{2^n\,2}+\mathbf{h}^{2^n\,3}+\mathbf{h}^{2^n\,4}+\mathbf{h}^{2^n\,5}+\ldots+\infty\right)\\ &\quad (\text{ by using } \mathbf{h}>0\ )\\ &= \frac{\mathbf{a}}{\mathbf{h}}\,\frac{\mathbf{h}^{2^n}}{1-\mathbf{h}^{2^n}}=\mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}}\\ &\quad (\text{ by using (12.52), namely } 0<\mathbf{h}<1\ )\end{aligned}$$

and that precisely leads us to

$$\left\|\overrightarrow{\mathbf{x}}_{n+k}-\overrightarrow{\mathbf{x}}_{n}\right\|<\mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}} \tag{12.65}$$

Here, for any arbitrarily chosen small positive number ϵ, there exists a natural number $N(\epsilon)$, such that

$$\mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}}<\epsilon \text{ for every } n>N(\epsilon) \tag{12.66}$$

and in fact, by

$$\mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}}<\epsilon \Leftrightarrow n>\log_2\left(\frac{\log\left(\frac{\mathbf{a}}{\mathbf{h}\epsilon}+1\right)}{\log\left(\frac{1}{\mathbf{h}}\right)}\right) \tag{12.67}$$

such that $\mathbf{a}>0$ and $0<\mathbf{h}<1$ are kept in mind and thereby the **positivity** of both the expressions $\log\left(\frac{\mathbf{a}}{\mathbf{h}\epsilon}+1\right)$ and $\log\left(\frac{1}{\mathbf{h}}\right)$ are ensured, the choice of $N(\epsilon)$ can be easily made by

$$N(\epsilon)=\left[\log_2\left(\frac{\log\left(\frac{\mathbf{a}}{\mathbf{h}\epsilon}+1\right)}{\log\left(\frac{1}{\mathbf{h}}\right)}\right)\right]+1 \tag{12.68}$$

where the above expression $\left[\log_2\left(\frac{\log\left(\frac{\mathbf{a}}{\mathbf{h}\epsilon}+1\right)}{\log\left(\frac{1}{\mathbf{h}}\right)}\right)\right]$ is the **integral part** of the expression $\log_2\left(\frac{\log\left(\frac{\mathbf{a}}{\mathbf{h}\epsilon}+1\right)}{\log\left(\frac{1}{\mathbf{h}}\right)}\right)$.

Therefore by combining (12.65) and (12.66), we get

$$\left\|\overrightarrow{\mathbf{x}}_{n+k}-\overrightarrow{\mathbf{x}}_{n}\right\|<\epsilon \text{ for every } n>N(\epsilon) \text{ and for every } k\in\mathbb{N}_0 \tag{12.69}$$

where the natural number $N(\epsilon)$ is given by (12.68).

Thus, by the **necessary and sufficient Cauchy's convergence criterion** for a sequence, the sequence $\{\overrightarrow{\mathbf{x}}_n\}_{n\in\mathbb{N}_0}$ **converges** and converges to a point, say $\overrightarrow{\xi}$ and this means

$$\lim_{k\to\infty} \overrightarrow{\mathbf{x}}_k = \overrightarrow{\xi} \tag{12.70}$$

So, by combining (12.64) and (12.70), which say that $\overrightarrow{\mathbf{x}}_{k+1} \in S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0)$ for every $k \in \mathbb{N}_0$ and $\lim_{k\to\infty} \overrightarrow{\mathbf{x}}_k = \overrightarrow{\xi}$ respectively, it is conclusively clear that

$$\overrightarrow{\xi} \in \overline{S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0)} \tag{12.71}$$

Again, by taking $k \to \infty$ on both the sides of (12.65), we get

$$\begin{aligned} &\lim_{k\to\infty} \left\| \overrightarrow{\mathbf{x}}_{n+k} - \overrightarrow{\mathbf{x}}_n \right\| \leq \lim_{k\to\infty} \mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}} = \mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}} \\ \Rightarrow &\left\| \overrightarrow{\xi} - \overrightarrow{\mathbf{x}}_n \right\| \leq \mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}} \text{ (by using (12.70))} \\ \Leftrightarrow &\left\| \overrightarrow{\mathbf{x}}_n - \overrightarrow{\xi} \right\| \leq \mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}} \end{aligned} \tag{12.72}$$

and thereby proving the statement (**C**) fully, namely (12.59).

Notably, the statement (**C**) stated by (12.59) or equivalently by (12.72) and the very fact $\mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}} < \epsilon$ for every $n > N(\epsilon)$ stated by (12.66) do **reconfirm the convergence** of the sequence $\{\overrightarrow{\mathbf{x}}_n\}_{n\in\mathbb{N}_0}$ to the point $\overrightarrow{\xi}$, because

$$\left\| \overrightarrow{\mathbf{x}}_n - \overrightarrow{\xi} \right\| \leq \mathbf{a}\frac{\mathbf{h}^{2^n-1}}{1-\mathbf{h}^{2^n}} < \epsilon \text{ for every } n > N(\epsilon)$$

In our final step, we need to show that $\overrightarrow{\xi}$ is a **zero** of the function $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})$, i.e. $\overrightarrow{\mathbf{f}}(\overrightarrow{\xi}) = \overrightarrow{0}$.

In order to show this, at first we restate that the statement (12.64), namely the statement $\overrightarrow{\mathbf{x}}_k \in S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0)$ for every $k \in \mathbb{N}_0$ implies that

$$\left\| \overrightarrow{\mathbf{x}}_k - \overrightarrow{\mathbf{x}}_0 \right\| < \mathbf{r} \text{ for every } k \in \mathbb{N}_0 \tag{12.73}$$

In that case, with the help of the established **lemma**, i.e. the duly proven statement (12.54) with regard to $\overrightarrow{\mathbf{x}}_0 \in S_{\mathbf{r}}(\overrightarrow{\mathbf{x}}_0)$ and $\left\| \overrightarrow{\mathbf{x}}_k - \overrightarrow{\mathbf{x}}_0 \right\| < \mathbf{r}$ for every $k \in \mathbb{N}_0$ (i.e. the very statement (12.73)), we get

$$\begin{aligned}
\left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0) \right\| &= \left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0 + \overrightarrow{\mathbf{s}}_k) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0) \right\| \\
&\leq \mathbf{c} \left\| \overrightarrow{\mathbf{s}}_k \right\| = \mathbf{c} \left\| \overrightarrow{\mathbf{x}}_k - \overrightarrow{\mathbf{x}}_0 \right\| < \mathbf{c}\,\mathbf{r} \\
&\qquad (\text{ by putting } \overrightarrow{\mathbf{s}}_k = \overrightarrow{\mathbf{x}}_k - \overrightarrow{\mathbf{x}}_0 \,) \\
\Rightarrow \left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k) \right\| - \left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0) \right\| &\leq \left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0) \right\| \leq \mathbf{c}\,\mathbf{r} \\
\Rightarrow \left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k) \right\| &\leq \left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0) \right\| + \mathbf{c}\,\mathbf{r} = \text{ a constant } = \mathbf{K} \text{ (say)}
\end{aligned} \tag{12.74}$$

Therefore, by using (12.60), namely by using $\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k)(\overrightarrow{\mathbf{x}}_{k+1} - \overrightarrow{\mathbf{x}}_k) = -\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k)$, we get

$$\begin{aligned}
\left\| \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k) \right\| = \left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k)(\overrightarrow{\mathbf{x}}_{k+1} - \overrightarrow{\mathbf{x}}_k) \right\| &\leq \left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k) \right\| \left\| \overrightarrow{\mathbf{x}}_{k+1} - \overrightarrow{\mathbf{x}}_k \right\| \\
&\leq \mathbf{K} \left\| \overrightarrow{\mathbf{x}}_{k+1} - \overrightarrow{\mathbf{x}}_k \right\| \\
&\qquad (\text{ by using (12.74) })
\end{aligned} \tag{12.75}$$

and therefore by taking limits on both the sides of (12.75) for $k \to \infty$, we finally arrive at

$$\begin{aligned}
\lim_{k\to\infty} \left\| \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k) \right\| &\leq \mathbf{K} \lim_{k\to\infty} \left\| \overrightarrow{\mathbf{x}}_{k+1} - \overrightarrow{\mathbf{x}}_k \right\| \\
\Rightarrow \left\| \overrightarrow{\mathbf{f}}(\overrightarrow{\xi}) \right\| &\leq \mathbf{K} \left\| \overrightarrow{\xi} - \overrightarrow{\xi} \right\| = 0 \\
&\qquad (\text{ by using (12.70) }) \\
\Rightarrow \overrightarrow{\mathbf{f}}(\overrightarrow{\xi}) &= \overrightarrow{0}
\end{aligned} \tag{12.76}$$

which eventually proves that $\overrightarrow{\xi}$ is a **zero** of the function $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})$.

Hence, (12.70), (12.71) and (12.76) are established and thereby proving the statement (**B**), namely (12.58). **This completes the proof of the convergence criterion of the Newton Raphson procedure**. □

An important remark: In the context of the $m \times m$ system of simultaneous equations (4.36), it has been shown that $\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})$ happens to be a **positive definite matrix** denoted by $\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}) = \mathsf{Cov}[\mathbf{X}_{(m)}]$ described by (4.39). $\mathsf{Cov}[\mathbf{X}_{(m)}]$ is called the covariance matrix of the m dimensional random vector $\mathbf{X}_{(m)} = (X, X^2, \ldots, X^m)$ (referred to (4.40)). The positive definiteness

of $\mathsf{Cov}[\mathbf{X}_{(m)}]$ has been **already proved** in the subsection 4.5.3 referred to the **positive definiteness of the covariance matrix of** $\mathbf{X}_{(m)}$. Thus, $\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})\right)^{-1}$ **exists for every** $\overrightarrow{\mathbf{x}} \in \mathbb{R}^m$ and hence the **convergence criterion of the Newton Raphson procedure is generally applicable in case of the system of equations (4.36)**.

12.3.4 Role of the Newton Raphson's convergence criterion in special cases

In this subsection, we shall confine our discussions about the convergence of the Newton Raphson procedure to the systems of equations (11.37) and (11.58) only. Each of these systems (11.37) and (11.58) are to be solved for β and γ. In this very subsection, just for the sake of a better degree of clarity (as per conventionally used symbols), we shall use the symbols x and y to denote the unknowns of the systems of equations instead of β and γ respectively.

In the light of solving the systems of equations (11.37) or (11.58), it has to be unforgettably stated that (12.95) and (12.96) already give the definition of our function $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})$. However, as we have already mentioned, we shall use (x, y) instead of (β, γ) to define the function $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}) = \overrightarrow{\mathbf{f}}(x, y) = \overrightarrow{0}$ **in this very subsection only**, after having denoted $\overrightarrow{\mathbf{f}} = \begin{pmatrix} f \\ \\ g \end{pmatrix}$.

In order to do so, we shall additionally use the symbol t instead of x and therefore, we shall denote the support of the probability distribution of a discrete X as $\{t_1, t_2, \ldots, t_N\}$ instead of $\{x_1, x_2, \ldots, x_N\}$.

By keeping this in mind, we derive and thereafter enlist the partial derivatives of f and g of the first order as

$$f_x(x, y) = \begin{cases} (\mu_2 - \mu_1^2) \sum\limits_{j=1}^{N} e^{xt_j + yt_j^2} & : \quad X \text{ is discrete} \\ (\mu_2 - \mu_1^2) \int\limits_{0}^{1} e^{xt + yt^2} dt & : \quad X \text{ is continuous} \end{cases} \tag{12.77}$$

$$f_y(x,y) = \begin{cases} (\mu_3 - \mu_1\mu_2) \sum\limits_{j=1}^{N} e^{xt_j + yt_j^2} & : \quad X \text{ is discrete} \\ (\mu_3 - \mu_1\mu_2) \int\limits_0^1 e^{xt+yt^2} dt & : \quad X \text{ is continuous} \end{cases} \tag{12.78}$$

$$g_x(x,y) = \begin{cases} (\mu_3 - \mu_1\mu_2) \sum\limits_{j=1}^{N} e^{xt_j + yt_j^2} & : \quad X \text{ is discrete} \\ (\mu_3 - \mu_1\mu_2) \int\limits_0^1 e^{xt+yt^2} dt & : \quad X \text{ is continuous} \end{cases} \tag{12.79}$$

$$g_y(x,y) = \begin{cases} (\mu_4 - \mu_2^2) \sum\limits_{j=1}^{N} e^{xt_j + yt_j^2} & : \quad X \text{ is discrete} \\ (\mu_4 - \mu_2^2) \int\limits_0^1 e^{xt+yt^2} dt & : \quad X \text{ is continuous} \end{cases} \tag{12.80}$$

In our cases, $\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}) = \begin{pmatrix} f_x(x,y) & f_y(x,y) \\ g_x(x,y) & g_y(x,y) \end{pmatrix}$ is a proven **positive definite covariance matrix** for every $\overrightarrow{\mathbf{x}} = (x,y) \in \mathbf{D} = \mathbb{R}^2$ and therefore the inverse of the same, namely $\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})\right)^{-1}$, exists for every $\overrightarrow{\mathbf{x}} \in \mathbf{D}$ (the positive definiteness of the matrix $\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})$ for every $\overrightarrow{\mathbf{x}} \in \mathbf{D}$ is **implied and implied by** the well established positive definiteness of the symmetric covariance matrix (6.30), the strict positiveness of the principal minors of which were duly established by (6.31) and (6.32)).

Before we proceed to discuss the utility of this aforesaid stated theorem 5.3.2, we need to state beforehand that, in every beginning of the course of an execution of the program for solving the system (11.37) or (11.58), the following things are to be duly pictured:

1. The vector $\overrightarrow{\mathbf{x}}_0$ is computed preliminarily on the basis of the user given inputs of μ_1 and μ_2. The word **preliminarily** is referred to the very fact that the starting vector $\overrightarrow{\mathbf{x}}_0$ **may** or **may not** fulfill the sufficient condition for convergence of the Newton Raphson procedure.

2. The existing complexity of the programming work is **not restricted** to a **single** user given input (μ_1, μ_2) only, but the programmer has to make sure that the preliminary value of $\overrightarrow{\mathbf{x}}_0$ is computable for **almost** every (μ_1, μ_2) belonging to the input space

$I_{(\mu_1,\mu_2)} = \{(\mu_1, \mu_2) \,|\, 0 < \mu_1 < 1\,,\, \mu_1^2 < \mu_2 < \mu_1\}$

The word **almost** is referred to the **important cases** only, which do span more than 98 % of the input space $I_{(\mu_1,\mu_2)}$, as far as the stochastic point of view of the solutions of these aforesaid systems of equations is concerned. The **unimportant cases** are referred to the stochastically irrelevant cases, when μ_2 is **unrealistically** closer to μ_1 from left.

3. With this value of $\overrightarrow{\mathbf{x}}_0$, the preliminary values of the sizes $\mathbf{a} = \left\| \left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0) \right)^{-1} \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_0) \right\|$ and $\mathbf{b} = \left\| \left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0) \right)^{-1} \right\|$ are computable. Subsequently, the preliminary values of the sizes $\mathbf{c}$, $\mathbf{h}$, $\mathbf{r}$ are $\mathbf{D}_0$ are consecutively computable.

Therefore, the sizes $\mathbf{a}$, $\mathbf{b}$ are computable and subsequently the sizes $\mathbf{c}$, $\mathbf{h}$, $\mathbf{r}$ are $\mathbf{D}_0$ are computable after every stage of **improvement** of $\overrightarrow{\mathbf{x}}_0$ (*obviously referred to the systems (11.37) and (11.58)*). This **improvement** of $\overrightarrow{\mathbf{x}}_0$ is well examinable by whether $\left\| \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_0) \right\|$ has become actually **smaller** than its previous value or not. In other words, an improvement of $\overrightarrow{\mathbf{x}}_0$ must cause the quantity $\left\| \overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_0) \right\|$ to get closer to 0 to an extent.

The procedures for the improvement of $\overrightarrow{\mathbf{x}}_0$ is **well programmable** and the procedures for the computations of these six sizes, namely $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{h}$, $\mathbf{r}$ and $\mathbf{D}_0$, are **theoretically programmable** too.

These six sizes are **meant to be** duly computed **after each and every improvement stage** of $\overrightarrow{\mathbf{x}}_0$ and this computation has to be repeatedly continued till $\overrightarrow{\mathbf{x}}_0$ fulfills the convergence conditions for the Newton Raphson procedure stated in the aforesaid theorem 5.3.2 and this should be carried out in a way that $\mathbf{D}_0$ could be an appropriate **circular neighbourhood** containing the point $\overrightarrow{\xi}$, within which any point would fulfill the necessary as well as the sufficient convergence condition of the Newton Raphson procedure.

However, it has to be unavoidably stated that the execution of the computations of all these six sizes, after each and every improvement stage of $\overrightarrow{\mathbf{x}}_0$, is rather **time consuming** and is bound to **prolong the running time** of the entire program for solving (11.37) or (11.58) unnecessarily. Long running times of these software programs are principally undesirable and as matter of

fact, a long running time would worsen one of the basic quality characteristics of a software program.

One thing is quite apparent that an effective programming of the procedures for computations of these six sizes is unavoidably hampered by the **repeated usage** of the **time consuming** subroutines, namely

- the **summation procedure**, for the cases of large values of N (for $N > 2000$), in case X happens to be discrete (referred to the system (11.37))
- the **numerical integration procedure**, especially for the cases when μ_2 is chosen to close to μ_1^2 from right or to close to μ_1 from left, in case X happens to be continuous (referred to the system (11.58))

Therefore, we could well point out, that at every improvement stage of $\overrightarrow{\mathbf{x}}_0$, the computations of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ necessitate the computation of a 2×2 matrix of the form $\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}})$, each of the elements of which involve either (aforesaid) summation procedures or numerical integration procedures. As we have already seen, these procedures could be well **time consuming** in many cases.

Additionally, programming an appropriate procedure for the computation of $\mathbf{c}$ with subject to the fulfillment of

- $\left\| \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}} + \overrightarrow{\mathbf{h_x}}) - \overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}) \right\| \leq \mathbf{c}\|\overrightarrow{\mathbf{h_x}}\|$ for every $\overrightarrow{\mathbf{x}}, \overrightarrow{\mathbf{y}} \in \mathbf{D}_0$, with $\overrightarrow{\mathbf{y}} = \overrightarrow{\mathbf{x}} + \overrightarrow{\mathbf{h_x}}$
- as well as $\mathbf{h} = \frac{\mathbf{abc}}{2} < 1$

with regard to the systems (11.37) and (11.58) is rather **cumbersome**. In fact, about a given suitably chosen neighbourhood of the point $\overrightarrow{\mathbf{x}} = \overrightarrow{\mathbf{x}}_0 = (x_0, y_0)$, any suitable utilizable computation of $\mathbf{c}$ demands **additional computations** of at least either four time consuming numerical integrations or four time consuming summations in form of computations of $\overrightarrow{\mathbf{f}}'(x_0 + h, y_0)$, $\overrightarrow{\mathbf{f}}'(x_0 - h, y_0)$, $\overrightarrow{\mathbf{f}}'(x_0, y_0 + k)$ and $\overrightarrow{\mathbf{f}}'(x_0, y_0 - k)$, h and k being suitably chosen increments (or decrements) of x_0 and y_0 respectively.

Apart from this, let us view an important characteristic property of both the systems of equations (11.37) and (11.58),

- the rate of change of $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})$ with respect to $\overrightarrow{\mathbf{x}}$ is **inconsiderably small** in **most of the cases**. This means, $\overrightarrow{\mathbf{x}}$ needs to be changed **drastically** towards the direction of $\overrightarrow{\xi}$ for making only **inconsiderably small** changes of $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})$ towards the direction of $\overrightarrow{0}$.

  This is particularly the case, when

  - μ_2 is chosen close to μ_1^2 from left or close to μ_1 from right.
  - μ_1 is chosen close to 0 from right or close to 1 from left.

- only in **few cases**, the rate of change of $\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}})$ with respect to $\overrightarrow{\mathbf{x}}$ is **considerably high**, which is good.

This precisely means that the programmer has to handle the size of $\|\overrightarrow{\mathbf{x}}_{k+1} - \overrightarrow{\mathbf{x}}_k\| = \left\|\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_k)\right)^{-1}\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_k)\right\|$ with a good amount of care, so as to make sure that the appropriate computation of $\mathbf{a}$, with respect to the fulfilment of $\left\|\left(\overrightarrow{\mathbf{f}}'(\overrightarrow{\mathbf{x}}_0)\right)^{-1}\overrightarrow{\mathbf{f}}(\overrightarrow{\mathbf{x}}_0)\right\| \leq \mathbf{a}$, actually takes place. This basically means that the **number of stages of improvement of $\overrightarrow{\mathbf{x}}_0$ is quite high in most of the cases**, which in turn **implies** that the **number of times** that these six sizes $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{h}$, $\mathbf{r}$ and $\mathbf{D}_0$ are needed to be computed is **quite high** too.

Hence, we arrive at the conclusion that the programming of the procedures for the computation of the six sizes $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{h}$, $\mathbf{r}$ and $\mathbf{D}_0$ is basically not worthy at all, because

- the computations of these six sizes after each and every improvement stage of $\overrightarrow{\mathbf{x}}_0$ are time consuming and therefore prolong the **running time** of the entire program.

- the **number of times** that these computations are to be carried out, is high in most of the cases.

- the **cumbersomeness** of the programming, as we have seen, is not avoidable.

Therefore, with subject to the consideration of these aforesaid arising difficulties, the software programs designed to solve the systems (11.37) and

(11.58), we conclusively find that the programming of this necessary and sufficient condition for the convergence of the **undamped** Newton Raphson procedure is completely unworthy.

This existing programming problem is **effectively** and **elegantly** resolvable by an alternative approach consisting of broadly two steps, **without hampering the convergence behavior or any convergence characteristic** of the sequence of **improved** values of $\overrightarrow{\mathbf{x}}_0$. Thus, this is alternative approach is **in no way a bad idea**, as far as the **essential quality characteristics of a software program** is concerned. These two steps are given as follows:

1. **Step 1**: For the solution of an equation having a single unknown, namely (12.10), we have already seen, that the Newton's condition for the convergence is given by (12.32). The extension of the condition for the convergence in case of two unknowns can be given to be

$$\frac{|f(x_0,y_0)f_{xx}(x_0,y_0)|}{\{f_x(x_0,y_0)\}^2} < 1 \tag{12.81}$$

$$\frac{|f(x_0,y_0)f_{yy}(x_0,y_0)|}{\{f_y(x_0,y_0)\}^2} < 1 \tag{12.82}$$

$$\frac{|g(x_0,y_0)g_{xx}(x_0,y_0)|}{\{g_x(x_0,y_0)\}^2} < 1 \tag{12.83}$$

$$\frac{|g(x_0,y_0)g_{yy}(x_0,y_0)|}{\{g_y(x_0,y_0)\}^2} < 1 \tag{12.84}$$

   The simultaneous fulfillment of all the four conditions would certainly leads to the fulfillment of the necessary condition for the desired convergence of Newton Raphson procedure and ensures the **correctness of the direction of convergence** of the sequence of improved values of $\overrightarrow{\mathbf{x}}_0$.

   Importantly, because of

   - by (6.35), $Var[X] = \mu_2 - \mu_1^2 > 0$
   - by (6.36), $Var[X^2] = \mu_4 - \mu_2^2 > 0$ and
   - by (6.37), $\mu_3 - \mu_1\mu_2 > 0$ (as a consequence of the proven inequality (5.4))

we easily conclude that the partial derivatives of the first order, namely (12.77), (12.78), (12.79) and (12.80) are always positive and hence all the fractions belonging to the constraints (12.81), (12.82), (12.83) and (12.84) are **finitely computable** (i.e. **the denominators of these fractions do not vanish**). Hence, there are no mathematical difficulties in this regard.

Practical experience shows that the fulfillment of all the four conditions stated above may not be stringently necessary, at least as far as our programming work is concerned.

Let R be a two-dimensional region containing all the possible first approximated solutions (x_0, y_0) fulfilling all the four above relations. Therefore, the area of R describes the smallness of expressions $T_f(h, k)$ and $T_g(h, k)$.

2. **Step 2**: Only after the vector $\overrightarrow{\mathbf{x}}_0 = (x_0, y_0)$ has fulfilled the necessary conditions for the convergence, further intermediate procedures with regard to the further improvements of $\overrightarrow{\mathbf{x}}_0$ are programmed prior to the programming of the **damped Newton Raphson procedure**.

   Importantly, it has to be clearly mentioned that these aforesaid developed software programs can **always be improved to any desired degree of goodness**, as far as the most desired quality characteristics are concerned.

12.3.5 Iterative procedure for two special equation-systems

Here, we shall discuss about two important systems of simultaneous equations, which shall be relevant for our present numerical work in the subsequent subsections.

For this, we define the first two moments of a random variable X having the range of variability $[0,1]$ as functions of two real variables namely β and γ as follows

$$E[X] = \mu_1^{(\beta,\gamma)} = \begin{cases} \frac{\sum_{j=1}^{N} x_j e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} & : \quad X \text{ is discrete} \\ \frac{\int_0^1 x e^{\beta x + \gamma x^2} dx}{\int_0^1 e^{\beta x + \gamma x^2} dx} & : \quad X \text{ is continuous} \end{cases} \tag{12.85}$$

$$E[X^2] = \mu_2^{(\beta,\gamma)} = \begin{cases} \frac{\sum_{j=1}^{N} x_j^2 e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} & : \quad X \text{ is discrete} \\ \frac{\int_0^1 x^2 e^{\beta x + \gamma x^2} dx}{\int_0^1 e^{\beta x + \gamma x^2} dx} & : \quad X \text{ is continuous} \end{cases} \tag{12.86}$$

where the random structure of X is defined by

- $f_{X|\{d\}}(x_j) = \frac{e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}}$, $0 = x_1 < x_2 < \ldots < x_N = 1$,
  in case X is a discrete random variable
- $f_{X|\{d\}}(x) = \frac{e^{\beta x + \gamma x^2}}{\int_0^1 e^{\beta x + \gamma x^2} dx}$, $0 \leq x \leq 1$,
  in case X is a continuous random variable

Now, if μ_1^* and μ_2^* be the preassigned values of the first two moments of X, then our problem will be to solve the following system of simultaneous equations (for β and γ) for the purpose of determining the random structure function of X

$$\begin{cases} \mu_1^{(\beta,\gamma)} - \mu_1^* &= 0 \\ \mu_2^{(\beta,\gamma)} - \mu_2^* &= 0 \end{cases} \tag{12.87}$$

Let (β^*, γ^*) be the solution to the system (12.87). Here, we shall introduce an iterative procedure for getting close to the solution of the system (12.87), if not a complete solution is possible. However, our experience shows, that the iterative procedure

- is rather a time consuming procedure, especially when X is continuous where the running times for the numerically computed integrations are too long.
- is a tentative process. This means, it may or may not lead to the desired solution (β^*, γ^*), but helps the intermediately computed vector (β, γ) to get closer to the solution vector (β^*, γ^*) reasonably well.

This closeness is described by the smallness of an introduced factor named *e_distance*, which is defined as

$$e\_distance = \sqrt{(\mu_1^{(\beta,\gamma)} - \mu_1^*)^2 + (\mu_2^{(\beta,\gamma)} - \mu_2^*)^2} \tag{12.88}$$

This means, if *e_distance* gets smaller and smaller, then (β, γ) gets closer and closer to the solution (β^*, γ^*) gradually. In this regard, the iterative procedure is of immense help, though the speed of convergence in certain cases may not be high enough. For our numerical procedures, we hold (β, γ) for a solution, if $e\_distance < 10^{-16}$ or at least $e\_distance < 10^{-10}$.

Before we give the formal algorithm of the iterative procedure, besides *e_distance*, we would like to introduce two more procedures and a variable:

- $$SolveForBeta(\gamma, \beta_s) \tag{12.89}$$

  is the procedure, which determines the value of β as the solution of the equation $\mu_1^{(\beta,\gamma)} - \mu_1^* = 0$ for a fixed value of γ and β_s as the starting value. This solution can be achieved by means of the numerical procedure (12.36), the running time of which can be considerably reduced by the introduction of β_s.

- $$SolveForGamma(\beta, \gamma_s) \tag{12.90}$$

  is the procedure, which determines the value of γ as the solution of the equation $\mu_2^{(\beta,\gamma)} - \mu_2^* = 0$ for a fixed value of β and γ_s as the starting

value. This solution can be achieved by means of the numerical procedure (12.36), the running time of which can be considerably reduced by the introduction of γ_s.

- The iterative procedure consists of a finite number of iterative steps, where the vector (β, γ) changes it's value after each such step. This change in each step is expected to lead to the direction of (β^*, γ^*). For to ensure the correctness of this direction, a *boolean* variable named *becomes_smaller* has been introduced.

  The variable *becomes_smaller* can have only two values, namely *true* and $false$. *becomes_smaller* has the value *true* after the completion of an iterative step, only if *e_distance* does not become larger as a result of that iterative step, otherwise $false$.

Therefore, the algorithm for the iterative procedure is described as:

Let β_s and γ_s be the starting values of β and γ respectively. In case, these starting values are not available, they are assumed as zeros.

We initialize $\beta = \beta_s$, $\gamma = \gamma_s$, *becomes_smaller* =*true*,

Compute *e_distance_updated* = *e_distance* with respect to β and γ;

while $((e\_distance\_updated > \epsilon) \wedge becomes\_smaller)$ {
$\beta_{reserve} = \beta$; $\gamma_{reserve} = \gamma$;
$\beta = SolveForBeta(\gamma, \beta)$;
$\gamma = SolveForGamma(\beta, \gamma)$
e_distance_current = *e_distance* with respect to β and γ;

if $(e\_distance\_updated < e\_distance\_current)$ then {
$\beta = \beta_{reserve}$; $\gamma = \gamma_{reserve}$;
$becomes\_smaller = false$;
}
else *e_distance_updated* = *e_distance_current*
}

(12.91)

(β, γ) is the final output as a result of the iterative procedure, ϵ being a given preassigned positive number ranging between 10^{-16} and 10^{-10}.

It has to be noted, that the iterative procedure ceases to be executed further, if a wrong direction is detected by *becomes_smaller* at the point at which the while-loop is broken.

It can be very well seen, that depending on the nature of the random variable X (ie. discrete or continuous) there are two different systems of simultaneous equations (in β and γ) of the form (12.87). Therefore, the iterative procedures for solving (12.87) in both the cases (X is discrete or continuous) are needed to be handled differently but in a similar manner.

12.3.6 Conditions for the convergence in cases of the two special systems

The necessary conditions for convergence corresponding to the vector (β, γ) shall be derived with the aim, that the Newton Raphson procedure is in a position to use this (β, γ) for reaching the solution (β^*, γ^*) to the system (12.87).

By rewriting the expressions $\mu_1^{(\beta,\gamma)} - \mu_1^*$ and $\mu_2^{(\beta,\gamma)} - \mu_2^*$ in the following way

$$\mu_1^{(\beta,\gamma)} - \mu_1^* = \begin{cases} \dfrac{\sum_{j=1}^{N} (x_j - \mu_1^*) e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} & : \quad X \text{ is discrete} \\ \dfrac{\int_0^1 (x - \mu_1^*) e^{\beta x + \gamma x^2} dx}{\int_0^1 e^{\beta x + \gamma x^2} dx} & : \quad X \text{ is continuous} \end{cases} \tag{12.92}$$

$$\mu_2^{(\beta,\gamma)} - \mu_2^* = \begin{cases} \dfrac{\sum_{j=1}^{N} (x_j^2 - \mu_2^*) e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} & : \quad X \text{ is discrete} \\ \dfrac{\int_0^1 (x^2 - \mu_2^*) e^{\beta x + \gamma x^2} dx}{\int_0^1 e^{\beta x + \gamma x^2} dx} & : \quad X \text{ is continuous} \end{cases} \tag{12.93}$$

the equation system (12.87) is equivalent to the following equation system

$$\begin{cases} f(\beta, \gamma) &= 0 \\ g(\beta, \gamma) &= 0 \end{cases} \tag{12.94}$$

such that

$$f(\beta,\gamma)=\begin{cases}\sum\limits_{j=1}^{N}(x_j-\mu_1^*)e^{\beta x_j+\gamma x_j^2} & : \quad X \text{ is discrete}\\ \int\limits_0^1(x-\mu_1^*)e^{\beta x+\gamma x^2}dx & : \quad X \text{ is continuous}\end{cases} \tag{12.95}$$

and

$$g(\beta,\gamma)=\begin{cases}\sum\limits_{j=1}^{N}(x_j^2-\mu_2^*)e^{\beta x_j+\gamma x_j^2} & : \quad X \text{ is discrete}\\ \int\limits_0^1(x^2-\mu_2^*)e^{\beta x+\gamma x^2}dx & : \quad X \text{ is continuous}\end{cases} \tag{12.96}$$

Since the systems (12.87) and (12.94) are equivalent with respect to their existing common solution (β^*,γ^*), we shall make use of the functions $f(\beta,\gamma)$ and $g(\beta,\gamma)$ for deriving the conditions of convergence. For this, we shall enlist their following partial derivatives:

$$f_\beta(\beta,\gamma)=\begin{cases}\sum\limits_{j=1}^{N}x_j(x_j-\mu_1^*)e^{\beta x_j+\gamma x_j^2} & : \quad X \text{ is discrete}\\ \int\limits_0^1 x(x-\mu_1^*)e^{\beta x+\gamma x^2}dx & : \quad X \text{ is continuous}\end{cases} \tag{12.97}$$

$$f_{\beta\beta}(\beta,\gamma)=\begin{cases}\sum\limits_{j=1}^{N}x_j^2(x_j-\mu_1^*)e^{\beta x_j+\gamma x_j^2} & : \quad X \text{ is discrete}\\ \int\limits_0^1 x^2(x-\mu_1^*)e^{\beta x+\gamma x^2}dx & : \quad X \text{ is continuous}\end{cases} \tag{12.98}$$

$$g_\gamma(\beta,\gamma)=\begin{cases}\sum\limits_{j=1}^{N}x_j^2(x_j^2-\mu_2^*)e^{\beta x_j+\gamma x_j^2} & : \quad X \text{ is discrete}\\ \int\limits_0^1 x^2(x^2-\mu_2^*)e^{\beta x+\gamma x^2}dx & : \quad X \text{ is continuous}\end{cases} \tag{12.99}$$

$$g_{\gamma\gamma}(\beta,\gamma)=\begin{cases}\sum\limits_{j=1}^{N}x_j^4(x_j^2-\mu_2^*)e^{\beta x_j+\gamma x_j^2} & : \quad X \text{ is discrete}\\ \int\limits_0^1 x^4(x^2-\mu_2^*)e^{\beta x+\gamma x^2}dx & : \quad X \text{ is continuous}\end{cases} \tag{12.100}$$

Accordingly, as already discussed in the previous subsection, the necessary conditions for the desired convergence with the help of (12.95), (12.98), (12.97) and (12.96), (12.100), (12.99) are given as

$$\frac{|f(\beta,\gamma)f_{\beta\beta}(\beta,\gamma)|}{\{f_\beta(\beta,\gamma)\}^2} = \frac{\left|\left(\mu_1^{(\beta,\gamma)} - \mu_1^*\right)\left(\mu_3^{(\beta,\gamma)} - \mu_1^*\mu_2^{(\beta,\gamma)}\right)\right|}{\left(\mu_2^{(\beta,\gamma)} - \mu_1^*\mu_1^{(\beta,\gamma)}\right)^2} \quad < \quad \kappa \quad (12.101)$$

$$\frac{|g(\beta,\gamma)g_{\gamma\gamma}(\beta,\gamma)|}{\{g_\gamma(\beta,\gamma)\}^2} = \frac{\left|\left(\mu_2^{(\beta,\gamma)} - \mu_2^*\right)\left(\mu_6^{(\beta,\gamma)} - \mu_2^*\mu_4^{(\beta,\gamma)}\right)\right|}{\left(\mu_4^{(\beta,\gamma)} - \mu_2^*\mu_2^{(\beta,\gamma)}\right)^2} \quad < \quad \kappa \quad (12.102)$$

where

- κ is a chosen positive real number lying between zero and one. Closer is κ to zero, stricter is the convergence condition

- $\mu_n^{(\beta,\gamma)} = E[X^n]$ with subject to the given values of β and γ for every $n \in \mathbb{N}$. From the programming point of view, the two cases, namely $\beta+\gamma < 709$ and $\beta+\gamma \geq 709$ must be handled separately, since the real value e^{709} exceeds the allowable upper limit of the java variable *double*. In the language of Java, $\beta+\gamma \geq 709$ means an overflow. On the other hand, an underflow as a result of the execution of a program is taken as a harmless zero. Keeping this in mind, we get

  - If X is discrete, then

$$\mu_n^{(\beta,\gamma)} = \begin{cases} \dfrac{\sum_{j=1}^{N} x_j^n e^{\beta x_j+\gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j+\gamma x_j^2}} & : \quad \beta+\gamma < 709 \\ \dfrac{\sum_{j=1}^{N} x_j^n e^{\beta(x_j-1)+\gamma(x_j^2-1)}}{\sum_{j=1}^{N} e^{\beta(x_j-1)+\gamma(x_j^2-1)}} & : \quad \beta+\gamma \geq 709 \end{cases} \quad (12.103)$$

  - If X is continuous, then $\mu_n^{(\beta,\gamma)}$ can be computed recursively as discussed before. This case will be thoroughly handled in this very subsection itself.

Our experience shows, that the introduction of the other two convergence conditions as discussed in the previous subsection, are not really necessary from the convergence point of view. Moreover, the introduction of the other two conditions mean a lot more computational time which is absolutely unnecessary.

Before we draw this subsection of convergence conditions to a close, we would like to introduce one more important thing, which will be absolutely necessary for programming our numerical procedural work in cases when X is a continuous random variable. Since we know, that the numerical integration is a time consuming procedure, it should be used only when it is absolutely necessary. For this, we have taken the following steps to ensure, that the computation of $\mu_n^{(\beta,\gamma)}$ for a given value of n involves the numerical integration just once, in cases when $\gamma \neq 0$. This very problem has already been discussed in the chapter 7 before.

The different cases involving the computation of $\mu_n^{(\beta,\gamma)}$ are therefore enlisted as

1. **Case 1** $\gamma = 0$, $\beta = 0$:

$$\mu_n^{(\beta,\gamma)} = \frac{1}{n+1} \tag{12.104}$$

   In particular, $\mu_1^{(\beta,\gamma)} = \frac{1}{2}$ and $\mu_2^{(\beta,\gamma)} = \frac{1}{3}$

2. **Case 2** $\gamma = 0$, $\beta \neq 0$:

$$\mu_n^{(\beta,\gamma)} = \begin{cases} 1 + \frac{1}{e^{\beta}-1} - \frac{1}{\beta} & : \quad n = 1 \\ 1 + \frac{1}{e^{\beta}-1} - \frac{n}{\beta}\mu_{n-1}^{(\beta,\gamma)} & : \quad n > 1 \end{cases} \tag{12.105}$$

   In particular,

$$\mu_2^{(\beta,\gamma)} = 1 + \frac{1}{e^{\beta}-1} - \frac{2}{\beta}\mu_1^{(\beta,\gamma)} = \frac{2(e^{\beta}-1) - \beta(2-\beta)e^{\beta}}{\beta^2(e^{\beta}-1)} \tag{12.106}$$

3. **Case 3** $\gamma \neq 0$:

- $\beta + \gamma < 709$:
  By using (7.100), (7.102) and (7.104), we get

$$\mu_n^{(\beta,\gamma)} = \begin{cases} \frac{1}{2\gamma}\left(\frac{e^{\beta+\gamma}-1}{\int_0^1 e^{\beta x+\gamma x^2}dx}-\beta\right) & : \quad n=1 \\ \frac{1}{2\gamma}\left(\frac{e^{\beta+\gamma}}{\int_0^1 e^{\beta x+\gamma x^2}dx}-1-\beta\mu_1^{(\beta,\gamma)}\right) & : \quad n=2 \\ \frac{1}{2\gamma}\left(\frac{e^{\beta+\gamma}}{\int_0^1 e^{\beta x+\gamma x^2}dx}-\beta\mu_{n-1}^{(\beta,\gamma)}-(n-1)\mu_{n-2}^{(\beta,\gamma)}\right) & : \quad n>2 \end{cases} \tag{12.107}$$

  In particular, if $\gamma = -\beta$, we can easily derive the following from the above relations

$$\mu_1^{(\beta,\gamma)} = \frac{1}{2} \tag{12.108}$$

$$\mu_2^{(\beta,\gamma)} = -\frac{1}{2\beta}\left(\frac{1}{\int_0^1 e^{\beta x(1-x)}dx}-1\right)+\frac{1}{4} \tag{12.109}$$

- $\beta + \gamma \geq 709$:
  By using (7.101), (7.103) and (7.105), we get

$$\mu_n^{(\beta,\gamma)} = \begin{cases} \frac{1}{2\gamma}\left(\frac{1-e^{-(\beta+\gamma)}}{\int_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx}-\beta\right) & : \quad n=1 \\ \frac{1}{2\gamma}\left(\frac{1}{\int_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx}-1-\beta\mu_1^{(\beta,\gamma)}\right) & : \quad n=2 \\ \frac{1}{2\gamma}\left(\frac{1}{\int_0^1 e^{\beta(x-1)+\gamma(x^2-1)}dx}-\beta\mu_{n-1}^{(\beta,\gamma)}-(n-1)\mu_{n-2}^{(\beta,\gamma)}\right) & : \quad n>2 \end{cases} \tag{12.110}$$

An important remark:

It is absolutely important to note, that the computations of the moments $\mu_n^{(\beta,\gamma)}$ $(n \in \mathbb{N})$ of X (wherever necessary) require the usage of

- (12.103) in cases when X is discrete and
- – (12.107) or (12.110), if $\gamma \neq 0$
  – (12.105), if $\beta \neq 0$ but $\gamma = 0$
  – (12.104), if $\beta = \gamma = 0$

  in cases when X is continuous

This helps the systematic computations of

- *e_distance* (given by 12.88) at every procedural step of the iterative procedure (12.91)
- *e_distance* (given by 12.88), together with the values of h and k (given in (12.114)) at every procedural step of the Newton Raphson procedure (12.115). This will be discussed in details in the very next subsection

12.3.7 Newton Raphson procedure for the two special equation-systems

The usefulness of the iterative procedure lies in the fact, that it brings the intermediately computed vector (β, γ) closer to the solution (β^*, γ^*). The chance, that the output vector (β, γ) as a result of the iterative procedure really fulfills the convergence conditions (12.101) and (12.102), becomes undoubtedly bigger. However, iterative procedure cannot be used as the procedure for the solution of any system of equations unless and until the solution could be reached easily, because it is a rather time consuming procedure. The Newton Raphson procedure should be used as a the procedure for the solution of a system instead, since it's running time is optimal.

In addition to the iterative procedure, there are other procedures for improving the vector (β, γ) in terms of it's closeness to (β^*, γ^*), which shall be discussed in the subsequent subsections. Our java programs are designed in a way, that the stated convergence conditions are fulfilled in most of the

cases with the help of the such improving procedures before forwarding the vector (β, γ) for the Newton Raphson solution procedure.

Therefore, for the purpose of designing our general algorithm for the solution of the system (12.94), the formal application of the stated theory of the (damped) Newton Raphson procedure (12.45) or even (12.46) with subject to an available (β_0, γ_0) as the first approximated solution of the (12.94) in both discrete and continuous cases, is imperatively necessary.

For this, we shall have to have a close look at the general procedural step of Newton Raphson:

For every $(n+2)^{th}$ $(n \in \mathbb{N}_0)$ procedural step of Newton Raphson, the computation of the $(n+2)^{th}$ approximated solution of (12.94), as described in (12.45), requires the computation of increments (or decrements) of β and γ given by h and k respectively as described in (12.44). That is, h and k for the $(n+2)^{th}$ step can be found by solving the two following equations simultaneously:

$$\begin{aligned} hf_\beta(\beta,\gamma) + kf_\gamma(\beta,\gamma) &= -f(\beta,\gamma) \\ hg_\beta(\beta,\gamma) + kg_\gamma(\beta,\gamma) &= -g(\beta,\gamma) \end{aligned}$$

which gives

$$\begin{aligned} h &= \tfrac{f_\gamma(\beta,\gamma)g(\beta,\gamma) - f(\beta,\gamma)g_\gamma(\beta,\gamma)}{\Psi_{h,k}} \\ k &= \tfrac{f(\beta,\gamma)g_\beta(\beta,\gamma) - f_\beta(\beta,\gamma)g(\beta,\gamma)}{\Psi_{h,k}} \text{ such that} \\ \Psi_{h,k} &= f_\beta(\beta,\gamma)g_\gamma(\beta,\gamma) - f_\gamma(\beta,\gamma)g_\beta(\beta,\gamma) \end{aligned} \tag{12.111}$$

such that the expressions of $f(\beta,\gamma)$, $f_\beta(\beta,\gamma)$, $g(\beta,\gamma)$ and $g_\gamma(\beta,\gamma)$ are described in (12.95), (12.97), (12.96) and (12.99) respectively, together with

$$f_\gamma(\beta,\gamma) = \begin{cases} \sum_{j=1}^{N} x_j^2(x_j - \mu_1^*)e^{\beta x_j + \gamma x_j^2} & : \ X \text{ is discrete} \\ \int_0^1 x^2(x-\mu_1^*)e^{\beta x + \gamma x^2}dx & : \ X \text{ is continuous} \end{cases} \tag{12.112}$$

and

$$g_\beta(\beta,\gamma) = \begin{cases} \sum_{j=1}^{N} x_j(x_j^2 - \mu_2^*)e^{\beta x_j + \gamma x_j^2} & : \quad X \text{ is discrete} \\ \int_0^1 x(x^2 - \mu_2^*)e^{\beta x + \gamma x^2} dx & : \quad X \text{ is continuous} \end{cases} \tag{12.113}$$

Therefore, by (12.95), (12.96), (12.97), (12.112), (12.113) and (12.99) we get

$$\begin{aligned} h &= \frac{\left(\mu_3^{(\beta,\gamma)} - \mu_1^*\mu_2^{(\beta,\gamma)}\right)\left(\mu_2^{(\beta,\gamma)} - \mu_2^*\right) - \left(\mu_1^{(\beta,\gamma)} - \mu_1^*\right)\left(\mu_4^{(\beta,\gamma)} - \mu_2^*\mu_2^{(\beta,\gamma)}\right)}{\Psi} \\ k &= \frac{\left(\mu_1^{(\beta,\gamma)} - \mu_1^*\right)\left(\mu_3^{(\beta,\gamma)} - \mu_2^*\mu_1^{(\beta,\gamma)}\right) - \left(\mu_2^{(\beta,\gamma)} - \mu_1^*\mu_1^{(\beta,\gamma)}\right)\left(\mu_2^{(\beta,\gamma)} - \mu_2^*\right)}{\Psi} \text{ such that} \\ \Psi &= \left(\mu_2^{(\beta,\gamma)} - \mu_1^*\mu_1^{(\beta,\gamma)}\right)\left(\mu_4^{(\beta,\gamma)} - \mu_2^*\mu_2^{(\beta,\gamma)}\right) - \left(\mu_3^{(\beta,\gamma)} - \mu_1^*\mu_2^{(\beta,\gamma)}\right)\left(\mu_3^{(\beta,\gamma)} - \mu_2^*\mu_1^{(\beta,\gamma)}\right) \end{aligned} \tag{12.114}$$

Hence, at each procedural step of Newton Raphson, (12.114) gives a clear picture about the increments (or decrements) of the values of β and γ, which are h and k respectively. Quite obviously, the values of h and k are said to be increments or decrements, according as they are positive or negative. If zero, then the corresponding value of β or γ is said to be unchanged.

The changed values of β and γ, which are $\beta + h$ and $\gamma + k$ respectively as a result of a given procedural step, take a step ahead in the direction of the aimed solution (β^*, γ^*).

However, importantly, the aforesaid addition of h and k to the values of β and γ respectively may mislead the direction of convergence to (β^*, γ^*), i.e the value of *e_distance* may be largened as a result of these additions of h and k. As a remedy, h and k can be **damped** by the factor t, $0 < t < 1$ before being added to β and γ respectively.

At first, each procedural step is started with $t = 1$. If *e_distance* gets largened in the process, then the (same) procedural step is repeated with t being halved, i.e. with $t = \frac{1}{2}$, after which it is examined by the program whether *e_distance* really gets smaller. If not, then t is halved again and this process is repeated. In this way, this process is repeated, till *e_distance* is really smaller. This is precisely the idea of damping the Newton Raphson procedure by the factor t.

Having this, the formal algorithm for the **(damped) Newton Raphson**

numerical procedure for the solution of (12.94), with subject to it's first approximated solution (β_0, γ_0), is developed as follows:

We initialize $\beta = \beta_0$, $\gamma = \gamma_0$; Select an ϵ, such that $\epsilon \in [10^{-16}, 10^{-10}]$;

Compute $\mu_1^{(\beta,\gamma)}$ and $\mu_2^{(\beta,\gamma)}$ with respect to β and γ;

$e\_distance = \sqrt{\left(\mu_1^{(\beta,\gamma)} - \mu_1^*\right)^2 + \left(\mu_2^{(\beta,\gamma)} - \mu_2^*\right)^2}$ as described in (12.88);

while $\{(\ e\_distance > \epsilon\ ) \wedge (additional\ criteria)\}$

{

Compute $\mu_3^{(\beta,\gamma)}$ and $\mu_4^{(\beta,\gamma)}$ with respect to β and γ;

$\hat{f} = \mu_1^{(\beta,\gamma)} - \mu_1^*$;

$\hat{f}_\beta = \mu_2^{(\beta,\gamma)} - \mu_1^*\mu_1^{(\beta,\gamma)}$; $\quad \hat{f}_\gamma = \mu_3^{(\beta,\gamma)} - \mu_1^*\mu_2^{(\beta,\gamma)}$;

$\hat{g} = \mu_2^{(\beta,\gamma)} - \mu_2^*$;

$\hat{g}_\beta = \mu_3^{(\beta,\gamma)} - \mu_2^*\mu_1^{(\beta,\gamma)}$; $\quad \hat{g}_\gamma = \mu_4^{(\beta,\gamma)} - \mu_2^*\mu_2^{(\beta,\gamma)}$;

$\Psi = \hat{f}_\beta\hat{g}_\gamma - \hat{f}_\gamma\hat{g}_\beta$;

Compute h and k according to (12.114), i.e $h = t\frac{\hat{f}_\gamma\hat{g} - \hat{f}\hat{g}_\gamma}{\Psi}$ & $k = t\frac{\hat{f}\hat{g}_\beta - \hat{f}_\beta\hat{g}}{\Psi}$;

β+= h;

γ+= k;

Compute $\mu_1^{(\beta,\gamma)}$ and $\mu_2^{(\beta,\gamma)}$ with respect to β and γ;

$e\_distance = \sqrt{\left(\mu_1^{(\beta,\gamma)} - \mu_1^*\right)^2 + \left(\mu_2^{(\beta,\gamma)} - \mu_2^*\right)^2}$

}

(β, γ) is the **final solution** on breaking of the above while-loop (12.115)

Notably, at any stage of the execution of the Newton Raphson procedure, if the condition *additional criteria* is not fulfilled, then the execution of Newton Raphson procedure is halted **prematurely**. The need of this *additional criteria* is left to the judgement of the programmer, otherwise it can also be omitted if not necessary.

12.4 Determination of a monotonic probability distribution

12.4.1 Algorithmic steps

A point of the ignorance space $\mathcal{D}_Y$ is represented as $d_Y = (\mu_Y^{(1)})$. The range of variability $\mathcal{X}_Y(\{d_Y\})$ of the random variable $Y|\{d_Y\}$ is a finite set of discrete real values. With this, the necessary steps for the computation of the probability distribution are:

1. **Input of data and check of compatibility:**

   - Cardinality of $\mathcal{X}_Y(\{d_Y\}) = |\mathcal{X}_Y(\{d_Y\})| = N$. Make sure, that $N \in \mathbb{N} \setminus \{0\}$
   - $\mathcal{X}_Y(\{d_Y\}) = \{y_1, y_2, \ldots, y_N\}$. Make sure, that all these y_js are distinct. Then, if not already sorted, then rearrange them, so that

   $$a = \min_{\mathcal{X}_Y(\{d_Y\})} y_j = y_1 < y_2 < \ldots < y_N = \max_{\mathcal{X}_Y(\{d_Y\})} y_j = b$$

   - Only if $N \geq 2$, then input Mean $= \mu_Y^{(1)}$ and make sure that $y_1 < \mu_Y^{(1)} < y_N$

2. **Linear transformation for $N \geq 2$ only:** $x_i = \frac{y_i - a}{b-a}$, such that

   - $f_{X|\{d\}}(x_j) = f_{Y|\{d_Y\}}(y_j)$
   - $f_{X|\{d\}}(x_j) = \frac{e^{\beta x_j}}{\sum_{j=1}^{N} e^{\beta x_j}}$
   - $f_{Y|\{d_Y\}}(y_j) = \frac{e^{\frac{\beta y_j}{b-a}}}{\sum_{j=1}^{N} e^{\frac{\beta y_j}{b-a}}} = f_{X|\{d\}}(x_j)$
   - $0 = x_1 < x_2 < x_3 < \ldots < x_N = 1$
   - $\mu_1^* = \frac{\mu_Y^{(1)} - a}{b-a}$

3. **Execution for $N \in \{1, 2\}$ only**: These are the special cases, which are handled without any numerical methods:

   (a) **Subcase 1**: $N = 1$

   This is the case for a degenerated probability distribution, where the inputs of both Mean and Variance are not allowable. In this case, we simply have $f_{Y|\{d_Y\}}(y_1) = 1$.

   (b) **Subcase 2**: $N = 2$

   This is the case for a two- point probability distribution, where the the input of Variance is not allowable. In this case, $a = y_1$ and $b = y_2$ and hence

   $$f_{Y|\{d_Y\}}(y_j) = (1 - \mu_1^*)e^{-\left(\log\left(\frac{1}{\mu_1^*}-1\right)\right)\left(\frac{y_j-a}{b-a}\right)} \quad (j = 1, 2)$$

4. **Numerical execution for $N > 2$ only:** Solve for β, the following equation numerically:

   $$\mu_1^* = \frac{\sum_{j=1}^{N} x_j e^{\beta x_j}}{\sum_{j=1}^{N} e^{\beta x_j}}$$

   The above equation can be rewritten as

   $$f(\beta) = \frac{\sum_{j=1}^{N} x_j e^{\beta x_j}}{\sum_{j=1}^{N} e^{\beta x_j}} - \mu_1^* = 0 \tag{12.116}$$

   which says, that $f(\beta)$ is continuous for all real values of β and therefore can be easily solved by the procedure (12.36) described before.

   Let β^* be the solution of $f(\beta) = 0$. Thus, $\lambda(d_Y) = (\lambda_1(d_Y)) = \left(\frac{\beta^*}{b-a}\right)$

5. **Final result for the user:**

$$f_{Y|\{d_Y\}}(y_j) = \begin{cases} 1 & : \quad N = 1 \\ (1-\mu_1^*)e^{-\left(\log\left(\frac{1}{\mu_1^*}-1\right)\right)\left(\frac{y_j-a}{b-a}\right)} & : \quad N = 2 \\ \frac{1}{N} & : \quad \mu_1^* = \frac{\sum\limits_{j=1}^{N} x_j}{N}, N > 2 \\ \frac{e^{\beta^*\left(\frac{y_j}{b-a}\right)}}{\sum\limits_{j=1}^{N} e^{\beta^*\left(\frac{y_j}{b-a}\right)}} & : \quad \mu_1^* \neq \frac{\sum\limits_{j=1}^{N} x_j}{N}, N > 2 \end{cases}$$

$$j = 1, 2, \ldots, N$$

12.4.2 The subfamily

Corresponding to the case, when $\mu_1^* = \frac{\sum\limits_{j=1}^{N} x_j}{N}$, we would arrive at the following discrete uniform distribution of Y described by

$$f_{Y|\{d_Y\}}(y_j) = \frac{1}{N}, \qquad j = 1, 2, \ldots, N$$

12.5 Determination of a continuous monotonic probability distribution

12.5.1 Algorithmic steps

A point of the ignorance space $\mathcal{D}_Y$ is represented as $d_Y = (\mu_Y^{(1)})$. The range of variability $\mathcal{X}_Y(\{d_Y\})$ of the random variable $Y|\{d_Y\}$ is a closed and bounded real interval. With this, the necessary steps for the computation of the probability distribution are:

1. **Input of data and it's compatibility:**
   - $\mathcal{X}_Y(\{d_Y\}) = \{y|a \leq y \leq b\}$. Make sure, that $a < b$, so that $|\mathcal{X}_Y(\{d_Y\})| = b - a$
   - Mean $= \mu_Y^{(1)}$. Make sure, that $a < \mu_Y^{(1)} < b$
2. **Linear transformation:** $x = \frac{y-a}{b-a}$, such that
   - $f_{X|\{d\}}(x) = (b-a) f_{Y|\{d_Y\}}(y) = \frac{e^{\beta x}}{\int_0^1 e^{\beta x} dx}$
   - $x \in [0, 1]$
   - $\mu_1^* = \frac{\mu_Y^{(1)} - a}{b-a}$
3. **Execution:** Solve for β the following equation:

$$\mu_1^* = 1 + \frac{1}{e^\beta - 1} - \frac{1}{\beta} \tag{12.117}$$

The above equation can be rewritten as

$$f(\beta) = 1 + \frac{1}{e^\beta - 1} - \frac{1}{\beta} - \mu_1^* = 0 \tag{12.118}$$

which says, that $f(\beta)$ is continuous for all nonzero real values of β, but is removable discontinuous at $\beta = 0$, which means

$$\lim_{\beta \to 0-} \left(1 + \frac{1}{e^\beta - 1} - \frac{1}{\beta} - \mu_1^*\right) = \lim_{\beta \to 0+} \left(1 + \frac{1}{e^\beta - 1} - \frac{1}{\beta} - \mu_1^*\right) = 0.5 - \mu_1^* \tag{12.119}$$

It has to be noted, that the right hand side of the relation (12.117) has been derived as the expression of $E[X]$ which has been shown in (12.105).

The relation (12.119) describes the removable discontinuity of $f(\beta)$ at the point $\beta = 0$. On removal of this discontinuity, the function $f(\beta)$ is redefined to make it a continuous function as follows:

$$f(\beta) = 0 = \begin{cases} 0.5 - \mu_1^* & : \quad \beta = 0 \\ 1 + \frac{1}{e^{\beta}-1} - \frac{1}{\beta} - \mu_1^* & : \quad \beta \neq 0 \end{cases} \tag{12.120}$$

which makes sure, that $f(\beta) = 0$ can be easily solved by the numerical procedure (12.36). As a result, the first moment of X denoted by $\mu_1 = \frac{\int_0^1 x e^{\beta x} dx}{\int_0^1 e^{\beta x} dx} = 1 + \frac{1}{e^{\beta}-1} - \frac{1}{\beta}$ can be plotted against β as follows:

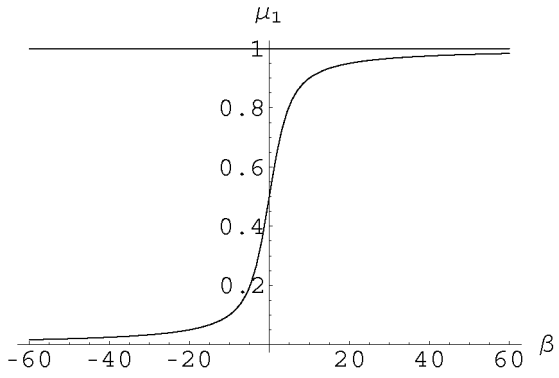


Figure 12.1: μ_1 against β

Let β^* be the solution of the equation $f(\beta) = 0$.

Therefore, $\lambda(d_Y) = (\lambda_1(d_Y)) = \left(\frac{\beta^*}{b-a}\right)$

4. **Final result for the user:**

$$f_{Y|\{d_Y\}}(y) = \begin{cases} \frac{\beta^* e^{\left(\frac{\beta^*}{b-a}\right) y}}{(b-a)\left(e^{\beta^*}-1\right) e^{\frac{a\beta^*}{b-a}}} & : \quad \mu_1^* \neq 0.5 \\ \frac{1}{b-a} & : \quad \mu_1^* = 0.5 \end{cases} \tag{12.121}$$

$$a \leq y \leq b$$

12.5.2 The subfamily

Corresponding to the case, when $\mu_1^* = 0.5$, we have

$$f_{Y|\{d_Y\}}(y) = \frac{1}{b-a}, \qquad a \leq y \leq b$$

which is a continuous uniform distribution of Y.

12.6 Determination of a discrete uni-extremal probability distribution

12.6.1 Algorithmic steps

A point of the ignorance space $\mathcal{D}_Y$ is represented as $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$. The range of variability $\mathcal{X}_Y(\{d_Y\})$ of the random variable $Y|\{d_Y\}$ is a finite set of discrete real values. With this, the necessary steps for the computation of the probability distribution are:

1. **Input of data and it's compatibility:**

   - Cardinality of $\mathcal{X}_Y(\{d_Y\}) = |\mathcal{X}_Y(\{d_Y\})| = N$. Make sure, that $N \in \mathbb{N} \setminus \{0\}$.
   - $\mathcal{X}_Y(\{d_Y\})$= $\{y_1, y_2, \ldots, y_N\}$. Make sure, that all these y_js are distinct. Then, if not already sorted, then rearrange them, so that

   $$a = \min_{\mathcal{X}_Y(\{d_Y\})} y_j = y_1 < y_2 < \ldots < y_N = \max_{\mathcal{X}_Y(\{d_Y\})} y_j = b$$

   - Only if $N \geq 2$, then input Mean = $\mu_Y^{(1)}$ and make sure that $y_1 < \mu_Y^{(1)} < y_N$
   - Only if $N \geq 3$, then input Variance = σ_Y^2 and make sure that $0 < \sigma_Y^2 < (\mu_Y^{(1)} - a)(b - \mu_Y^{(1)})$.

     The user is given a choice to give the second moment $\mu_Y^{(2)}$ of the probability distribution instead of the variance $\sigma_Y^2 = \mu_Y^{(2)} - (\mu_Y^{(1)})^2$, if he wishes. In that case, the user has to make sure, that

     $\left(\mu_Y^{(1)}\right)^2 < \mu_Y^{(2)} < (a+b)\mu_Y^{(1)} - ab$ and then compute σ_Y^2

2. **Linear transformation for $N \geq 2$ only:** $x_i = \frac{y_i - a}{b-a}$, such that

   - $f_{X|\{d\}}(x_j) = f_{Y|\{d_Y\}}(y_j)$
   - $f_{X|\{d\}}(x_j) = \frac{e^{\beta x_j + \gamma x_j^2}}{\sum\limits_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}}$
   - $f_{Y|\{d_Y\}}(y_j) = \frac{e^{\beta\left(\frac{y_j - a}{b-a}\right) + \gamma\left(\frac{y_j - a}{b-a}\right)^2}}{\sum\limits_{j=1}^{N} e^{\beta\left(\frac{y_j - a}{b-a}\right) + \gamma\left(\frac{y_j - a}{b-a}\right)^2}} = f_{X|\{d\}}(x_j)$
   - $0 = x_1 < x_2 < x_3 < \ldots < x_N = 1$
   - $\mu_1^* = \frac{\mu_Y^{(1)} - a}{b-a}$
   - $\mu_2^* = \mu_1^{*2} + \frac{\sigma_Y^2}{(b-a)^2}$

3. **Execution for $N \in \{1, 2, 3\}$ only**: These are the special cases, which are handled without any numerical methods:

   (a) **Subcase 1**: $N = 1$

   This is the case for a degenerated probability distribution, where the inputs of both Mean and Variance are not allowable. In this case, we simply have $f_{Y|\{d_Y\}}(y_1) = 1$.

   (b) **Subcase 2**: $N = 2$

   This is the case for a two- point probability distribution, where the input of the Mean is allowable, but the input of the Variance is not allowable. In this case, $a = y_1$ and $b = y_2$ and hence

   $$f_{Y|\{d_Y\}}(y_j) = (1 - \mu_1^*) e^{-\left(\log\left(\frac{1}{\mu_1^*} - 1\right)\right)\left(\frac{y_j - a}{b-a}\right)} \quad (j = 1, 2)$$

   (c) **Subcase 3**: $N = 3$

   This is the case for a three- point probability distribution, where the inputs of both Mean and Variance are allowable. In this case, $a = y_1$ and $b = y_3$. However, the input Variance cannot be chosen arbitrarily small.

   In this case, at first the following system of equations is solved for p_1, p_2 and p_3, by keeping $x_1 = 0, x_2 = \hat{x}, x_3 = 1$ in mind:

$$\begin{cases} 1 & = p_1 + p_2 + p_3 \\ \mu_1^* & = x_1 p_1 + x_2 p_2 + x_3 p_3 = \hat{x} p_2 + p_3 \\ \mu_2^* & = x_1^2 p_1 + x_2^2 p_2 + x_3^2 p_3 = \hat{x}^2 p_2 + p_3 \end{cases}$$

and only the feasible solution, i.e. $p_1 > 0, p_2 > 0$ and $p_3 > 0$ is taken. If a feasible solution does not exist, then μ_2^* is increased till the feasibility of the solution is marginally reached.

After this, the values of α^*, β^* and γ^* are computed as follows:

- $\alpha^* = \log(p_1)$
- $\beta^* = \frac{\log\left(\frac{p_2}{p_1}\right) - \hat{x}^2 \log\left(\frac{p_3}{p_1}\right)}{\hat{x}(1-\hat{x})}$
- $\gamma^* = \frac{\hat{x} \log\left(\frac{p_3}{p_1}\right) - \log\left(\frac{p_2}{p_1}\right)}{\hat{x}(1-\hat{x})}$

and thereby giving the required probability distribution as

$$f_{Y|\{d_Y\}} = e^{\alpha^* + \beta^* \left(\frac{y_j - a}{b-a}\right) + \gamma^* \left(\frac{y_j - a}{b-a}\right)^2}, \quad (j = 1, 2, 3)$$

4. **Numerical execution for $N > 3$ only**: Let (β^*, γ^*) be the solution of the following system of simultaneous equations in β and γ:

$$\begin{cases} \mu_1^* = \frac{\sum_{j=1}^{N} x_j e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} \\ \mu_2^* = \frac{\sum_{j=1}^{N} x_j^2 e^{\beta x_j + \gamma x_j^2}}{\sum_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} \end{cases} \tag{12.122}$$

which signifies $\lambda(d_Y) = (\lambda_1(d_Y), \lambda_2(d_Y)) = \left(\frac{\beta^*(b-a) - 2a\gamma^*}{(b-a)^2}, \frac{\gamma^*}{(b-a)^2}\right)$.

Now, for the purpose of getting (β^*, γ^*) with subject to the minimization of the running time of the program, we shall have to go through the following steps where different cases are handled sequentially, rather than designing a generalized algorithm for finding a solution to (12.122) without any consideration for the special cases:

(a) **case 1** (**Uniformity may hold**): If the following two relations hold

$$\mu_1^* = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{12.123}$$

$$\mu_2^* = \frac{1}{N}\sum_{j=1}^{N} x_j^2 \tag{12.124}$$

then $\beta^* = 0$ and $\gamma^* = 0$ and the probability distribution of Y is a discrete uniform distribution, whose probability mass function $f_{Y|\{d_Y\}}(y_j)$, $j = 1, 2, \ldots, N$ as the final result is given as

$$f_{Y|\{d_Y\}}(y_j) = \frac{1}{N} \quad (j = 1, 2, \ldots, N) \tag{12.125}$$

after which the further execution of the program has to be **stopped**. Otherwise, if this is not the case, then proceed to the next case

(b) **case 2** (**Monotonicity may hold**): At first, solve the following equation in β numerically

$$\mu_1^* = \frac{\sum_{j=1}^{N} x_j e^{\beta x_j}}{\sum_{j=1}^{N} e^{\beta x_j}}$$

and let β_0 be it's solution. This β_0 is determined by rewriting the above equation exactly in the same way as (12.116) and then solving it by the procedure (12.36) as described before. Then, if the following

$$\mu_2^* = \frac{\sum_{j=1}^{N} x_j^2 e^{\beta_0 x_j}}{\sum_{j=1}^{N} e^{\beta_0 x_j}} \tag{12.126}$$

holds true, then $\beta^* = \beta_0$ and $\gamma^* = 0$ and the probability distribution of Y is a discrete MEP- monotonic probability distribution, whose probability mass function $f_{Y|\{d_Y\}}(y_j)$, $j = 1, 2, \ldots, N$ as the

final result is given as

$$f_{Y|\{d_Y\}}(y_j) = \frac{e^{\beta_0\left(\frac{y_j}{b-a}\right)}}{\sum_{j=1}^{N} e^{\beta_0\left(\frac{y_j}{b-a}\right)}} \quad (j = 1, 2, \ldots, N) \tag{12.127}$$

after which the further execution of the program has to be **stopped**. Otherwise, if this is not the case, then proceed to the next case

(c) **case 3** (**Probability distribution with a small variance**): With subject to a given approximating condition, when the given variance σ_Y^2 is treated to be small enough with respect to the given value of mean $\mu_Y^{(1)}$, the probability distribution of Y is approximated to the truncated discrete normal distribution whose mean and variance are approximately $\mu_Y^{(1)}$ and σ_Y^2 respectively, the probability mass function being

$$f_{Y|\{d_Y\}}(y_j) = \frac{e^{-\frac{1}{2}\left(\frac{y_j-\mu_Y^{(1)}}{\sigma_Y}\right)^2}}{\sum_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{y_j-\mu_Y^{(1)}}{\sigma_Y}\right)^2}} \quad (j = 1, 2, \ldots, N) \tag{12.128}$$

If the approximating condition holds, then **stop** the further execution of the program at this point, giving the final result (12.128). Otherwise, proceed to the next case.

The details of achieving this approximated probability distribution are given in one of the following subsequent subsections.

(d) **case 4** (**None of the above cases**): Lastly, if none of the above is the case, then solve the system of the given simultaneous equations (12.122) numerically, which corresponds to a general discrete MEP- uni- extremal probability distribution.

For the purpose of finding the solution (β^*, γ^*) of (12.122) by the Newton- Raphson method (described in [36]), a suitable first approximated solution of (12.122) is absolutely necessary.

However, it has to be pointed out that the solution (β^*, γ^*) does not exist corresponding to every pair (μ_1^*, μ_2^*) in cases when μ_2^* is

too close to μ_1^{*2}. In such cases, the program **attempts to give** the solution (β^*, γ^*) corresponding to the smallest possible $\widetilde{\mu_2} \geq \mu_2^*$.

The details of solving this general case are discussed in one of the following subsequent subsections.

The final results of all the cases are summarized in the following step.

5. **Final result for the user:**

$$f_{Y|\{d_Y\}}(y_j)$$

$$= \begin{cases} 1 & : \quad N = 1 \\ (1-\mu_1^*)e^{-\left(\log\left(\frac{1}{\mu_1^*}-1\right)\right)\left(\frac{y_j-a}{b-a}\right)} & : \quad N = 2 \\ e^{\widetilde{\alpha}+\widetilde{\beta}\left(\frac{y_j-a}{b-a}\right)+\widetilde{\gamma}\left(\frac{y_j-a}{b-a}\right)^2} & : \quad N = 3 \\ \frac{1}{N} & : \quad \mu_1^* = \frac{\sum_{j=1}^{N} x_j}{N}, \mu_2^* = \frac{\sum_{j=1}^{N} x_j^2}{N}, N > 3 \\ \frac{e^{\beta^*\left(\frac{y_j}{b-a}\right)}}{\sum_{j=1}^{N} e^{\beta^*\left(\frac{y_j}{b-a}\right)}} & : \mu_1^*, \mu_2^* \text{ for monotonicity}, N > 3 \\ \frac{e^{-\frac{1}{2}\left(\frac{y_j-\mu_Y^{(1)}}{\sigma_Y}\right)^2}}{\sum_{j=1}^{N} e^{-\frac{1}{2}\left(\frac{y_j-\mu_Y^{(1)}}{\sigma_Y}\right)^2}} & : \text{for a small variance}, N > 3 \\ \frac{e^{\lambda_1^* y_j+\lambda_2^* y_j^2}}{\sum_{j=1}^{N} e^{\lambda_1^* y_j+\lambda_2^* y_j^2}} & : \quad N > 3, \textbf{the non special case} \end{cases}$$

$$j = 1, 2, \ldots, N$$

such that in the above stated non special case,

- $\lambda_1^* = \frac{\beta^*(b-a)-2a\gamma^*}{(b-a)^2}$
- $\lambda_2^* = \frac{\gamma^*}{(b-a)^2}$

12.6.2 The subroutine for non special uni-extremal cases

The solution of the system of equations (12.122), where X is a discrete random variable with the range of variability $\{x_1, x_2, \ldots, x_N\}$, is identical with the solution of the system of equations (12.94) for the discrete case. This justifies the usage of the designed Newton Raphson numerical procedure (12.115) for the solution of the system (12.122) as well.

Therefore, the non special case algorithm, which aims to solve the system (12.122) step by step, is described by the following sequential steps:

1. Solve the following equation (in β) by the procedure for the complete solution (12.36) after resetting the equation into $f(\beta) = 0$ form:

$$\mu_1^* = \frac{\sum_{j=1}^{N} x_j e^{\beta x_j}}{\sum_{j=1}^{N} e^{\beta x_j}}$$

   Let the yielded solution be β_s

2. By taking the starting vector (β_0, γ_0) for $\beta_0 = \beta_s$ and $\gamma_0 = 0$, the iterative procedure (12.91) is executed for $\epsilon = 10^{-10}$ in a way, that the simultaneous fulfillment of the convergence conditions described by (12.101) and (12.102) are examined for κ =0.5 at every iterative step.

   If these convergence conditions are fulfilled simultaneously at any such step, then the iterative procedure must be discontinued after that particular step immediately. Otherwise the iterative procedure must be continued as described.

   Let the final resulting vector be (β, γ) at the end of this iterative procedure

3. Using this (β, γ) as the first approximated solution of the system (12.94), apply the Newton Raphson procedure (12.115) in the following way:

   (a) *e_distance* (described in (12.88)) is computed before and after the execution of the $(n+2)^{th}$ Newton Raphson procedural step each time, such that $n \in \mathbb{N}_0$. Let these two *e_distance*s be named

as *e_distance_previous* and *e_distance_current* respectively. This $(n+2)^{th}$ step corresponds to the computation of the $(n+2)^{th}$ approximated solution by the Newton Raphson procedure. If $e\_distance\_current \leq e\_distance\_previous$ then the boolean variable $becomes\_smaller = true$, otherwise $becomes\_smaller = false$.

(b) The Newton Raphson procedure must be discontinued immediately, if at least one of the following conditions are violated:

- $becomes\_smaller = true$
- $e\_distance\_current \geq 10^{-12}$
- $n < 100000$

In this case, $\epsilon = 10^{-12}$. The simultaneous fulfillment of the conditions, namely $becomes\_smaller = true$ and $n < 100000$ is equivalent to the programmer-constructed condition *additional criteria* (as described in the procedure (12.115) previously).

Let the final resulting vector be (β, γ) at the end of this Newton Raphson procedure

4. If $e\_distance > 10^{-10}$ still happens to be true corresponding to (β, γ), then the iterative procedure (12.91) corresponding to the starting vector (β, γ) is executed once again for $\epsilon = 10^{-10}$ and for at most 100000 iterations.

   Let the final resulting vector be (β, γ) at the end of this iterative procedure

5. If $e\_distance > 10^{-10}$ still happens to be true corresponding to (β, γ), then by using this (β, γ) as the first approximated solution of the system (12.94), apply the Newton Raphson procedure (12.115) once again. The execution of this procedure is continued, till one of the following situations is reached:

   - $e\_distance < 10^{-10}$ is reached corresponding to the current values of β and γ at a particular procedural step
   - number of procedural steps reaches 10000

   In this case, $\epsilon = 10^{-12}$ and $n < 10000$ is taken to be the condition *additional criteria*.

Additionally, throughout the entire Newton Raphson procedure, that particular intermediate vector (β, γ) is taken as the final resulting vector, corresponding to which the *e_distance* is minimum. Let such (β, γ) be (β^*, γ^*).

At this juncture, there are two possibilities, namely

- *e_distance* $< 10^{-10}$ is reached: This means (β^*, γ^*) is the desired solution of (12.122).
- *e_distance* $\geq 10^{-10}$ holds still: This means (β^*, γ^*) is not the desired solution of (12.122) and in other words, the solution of (12.122) does not exist in this case.

  Evidently, (β^*, γ^*) is the solution of the following system:

$$\begin{cases} \mu_1^* = \frac{\sum\limits_{j=1}^{N} x_j e^{\beta x_j + \gamma x_j^2}}{\sum\limits_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} \\ \widetilde{\mu_2} = \frac{\sum\limits_{j=1}^{N} x_j^2 e^{\beta x_j + \gamma x_j^2}}{\sum\limits_{j=1}^{N} e^{\beta x_j + \gamma x_j^2}} \end{cases}$$

  such that $\widetilde{\mu_2} > \mu_2^*$ and $\widetilde{\mu_2}$ is a computed value of the second moment of X.

Thus, the computed (β^*, γ^*) [1] in the final step is either the desired solution of (12.122) or a possible utilizable value of the desired vector for the construction of our probability distribution of X.

Whence, by using (6.75) too, $f_{Y|\{d_Y\}}(y_j)$ is finally expressed as

$$f_{Y|\{d_Y\}}(y_j) = \frac{e^{\beta^*\left(\frac{y_j-a}{b-a}\right)+\gamma^*\left(\frac{y_j-a}{b-a}\right)^2}}{\sum\limits_{j=1}^{N} e^{\beta^*\left(\frac{y_j-a}{b-a}\right)+\gamma^*\left(\frac{y_j-a}{b-a}\right)^2}} = \frac{K}{\sigma_Y\sqrt{2\pi}} e^{\widehat{\lambda}\left(\frac{y_j-M}{\sigma_Y}\right)^2}, \quad j = 1, 2, \ldots, N \tag{12.129}$$

[1] The **difference** between $\widetilde{\mu_2}$ and μ_2^* says that the numerical solution of (12.122) **does not exist** and this is beyond programmer's control. However, $\widetilde{\mu_2}$ is **sometimes** the **smallest possible value** of the second moment of X

12.7 Determination of a continuous uni-extremal probability distribution

12.7.1 Algorithmic steps

A point of the ignorance space $\mathcal{D}_Y$ is represented as $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)})$. The range of variability $\mathcal{X}_Y(\{d_Y\})$ of the random variable $Y|\{d_Y\}$ is a closed and bounded real interval. With this, the necessary steps for the computation of the probability distribution are:

1. **Input of data and it's compatibility:**
   - $\mathcal{X}_Y(\{d_Y\}) = \{y|a \leq y \leq b\}$. Make sure, that $a < b$
   - Mean $= \mu_Y^{(1)}$. Make sure, that $a < \mu_Y^{(1)} < b$
   - Variance $= \sigma_Y^2$. Make sure, that $0 < \sigma_Y^2 < (\mu_Y^{(1)} - a)(b - \mu_Y^{(1)})$. The user is given a choice to give the second moment $(\mu_Y^{(2)})$ of the probability distribution instead of the variance (σ_Y^2), if he wishes. In that case, the user has to make sure, that $(\mu_Y^{(1)})^2 < \mu_Y^{(2)} < (a+b)\mu_Y^{(1)} - ab$ and only after that compute $\sigma_Y^2 = \mu_Y^{(2)} - (\mu_Y^{(1)})^2$
2. **Linear transformation:** $x = \frac{y-a}{b-a}$, such that
   - $f_{X|\{d\}}(x) = (b-a) f_{Y|\{d_Y\}}(y) = \frac{e^{\beta x + \gamma x^2}}{\int_0^1 e^{\beta x + \gamma x^2} dx}$
   - $x \in [0, 1]$
   - $\mu_1^* = \frac{\mu_Y^{(1)} - a}{b-a}$
   - $\mu_2^* = \mu_1^{*2} + \frac{\sigma_Y^2}{(b-a)^2}$
3. **Execution:** Let (β^*, γ^*) be the solution of the following system of simultaneous equations in β and γ:

$$\begin{cases} \mu_1^* = \dfrac{\int_0^1 x e^{\beta x + \gamma x^2} dx}{\int_0^1 e^{\beta x + \gamma x^2} dx} \\ \mu_2^* = \dfrac{\int_0^1 x^2 e^{\beta x + \gamma x^2} dx}{\int_0^1 e^{\beta x + \gamma x^2} dx} \end{cases} \tag{12.130}$$

which signifies $\lambda(d_Y) = (\lambda_1(d_Y), \lambda_2(d_Y)) = \left(\frac{\beta^*(b-a)-2a\gamma^*}{(b-a)^2}, \frac{\gamma^*}{(b-a)^2}\right)$.

Now, for the purpose of getting (β^*, γ^*) with subject to the minimization of the running time of the program, we shall have to go through the following steps where different cases are handled sequentially, rather than designing a generalized algorithm for finding a solution to (12.130) without any consideration for the special cases:

(a) **case 1** ($\mu_1^* = 0.5, \mu_2^* = \frac{1}{3}$): $\beta^* = 0$, $\gamma^* = 0$. The probability distribution of Y is the continuous uniform distribution, whose probability density function $f_{Y|\{d_Y\}}(y)$, $a \leq y \leq b$ as the final result is given as

$$f_{Y|\{d_Y\}}(y) = \frac{1}{b-a}, \qquad a \leq y \leq b \tag{12.131}$$

after which the further execution of the program has to be **stopped**. Otherwise, if this is not the case, then proceed to the next case

(b) **case 2** ($\mu_1^* = 0.5, \mu_2^* \neq \frac{1}{3}$): $\beta^* = -\gamma^*$ is the solution of the following equation in β

$$\mu_2^* + \frac{1}{2\beta}\left(\frac{1}{\int\limits_0^1 e^{\beta t(1-t)}dt} - 1\right) = 0.25 \tag{12.132}$$

which is the result of the usage of (12.109). β^* can be easily determined by means of the numerical procedure (12.36). The probability distribution of Y is a symmetric continuous MEP-uni- extremal probability distribution, whose probability density function $f_{Y|\{d_Y\}}(y)$, $a \leq y \leq b$ as the final result is given as

$$f_{Y|\{d_Y\}}(y) = \frac{e^{\frac{\beta^*(y-a)(b-y)}{(b-a)^2}}}{(b-a)\int\limits_0^1 e^{\beta^* t(1-t)}dt}, \qquad a \leq y \leq b \tag{12.133}$$

after which the further execution of the program has to be **stopped**. Otherwise, if this is not the case, then proceed to the next case.

(c) **case 3** ($\mu_1^* \neq 0.5$ **and the monotonicity may hold**): At first, solve the following equation in β numerically

$$\mu_1^* = 1 + \frac{1}{e^{\beta} - 1} - \frac{1}{\beta}$$

and let β_0 be it's solution. This β_0 is determined exactly in the identical manner as the equation (12.117) was rewritten in the form of (12.120) and then solved by means of the procedure (12.36). Then, if the following holds true

$$\mu_2^* = \frac{2(e^{\beta_0} - 1) - \beta_0(2 - \beta_0)e^{\beta_0}}{\beta_0^2(e^{\beta_0} - 1)} \tag{12.134}$$

then by using (12.106), we conclude $\beta^* = \beta_0$ and $\gamma^* = 0$. The probability distribution of Y is a continuous MEP- monotonic probability distribution, whose probability density function $f_{Y|\{d_Y\}}(y)$, $a \leq y \leq b$ as the final result is given as

$$f_{Y|\{d_Y\}}(y) = \frac{\beta_0 e^{\left(\frac{\beta_0}{b-a}\right)y}}{(b-a)\left(e^{\beta_0} - 1\right)e^{\frac{a\beta_0}{b-a}}} \qquad a \leq y \leq b \tag{12.135}$$

after which the further execution of the program has to be **stopped**. Otherwise, if this is not the case, then proceed to the next case.

(d) **case 4** (**Probability distribution with a small variance**): With subject to a given approximating condition, when the given variance σ_Y^2 is treated to be small enough with respect to the given value of mean $\mu_Y^{(1)}$, the probability distribution of Y is approximated to the truncated normal distribution whose mean and variance are approximately $\mu_Y^{(1)}$ and σ_Y^2 respectively. By $\mu_1^* = \frac{\mu_Y^{(1)} - a}{b-a}$ and $\sigma = \frac{\sigma_Y}{b-a}$, the probability density function is given as

$$f_{Y|\{d_Y\}}(y) = \frac{1}{(b-a)\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1^*}{\sigma}\right)^2} dx} e^{-\frac{1}{2}\left(\frac{y-\mu_Y^{(1)}}{\sigma_Y}\right)^2}, \qquad a \leq y \leq b \tag{12.136}$$

If the approximating condition holds, then **stop** the further execution of the program at this point, giving the result (12.136). Otherwise, proceed to the next case.

The details of achieving this approximated probability distribution are given in one of the following subsequent subsections.

(e) **case 5** (**None of the above cases**): Lastly, if none of the above is the case, then solve the system of the given simultaneous equations (12.130) numerically, which corresponds to a general continuous MEP- uni- extremal probability distribution.

For the purpose of finding the solution (β^*, γ^*) of (12.130) by the Newton- Raphson method (described in [36]), a suitable first approximated solution of (12.130) is absolutely necessary.

The details of solving this general case are discussed in the subsequent subsections.

The final results of all the cases are summarized in the following step.

4. **Final result for the user:**

$$f_{Y|\{d_Y\}}(y) = \begin{cases} \frac{1}{b-a} & : \quad \mu_1^* = 0.5, \mu_2^* = \frac{1}{3} \\ \frac{e^{\frac{\beta^*(y-a)(b-y)}{(b-a)^2}}}{(b-a)\int\limits_0^1 e^{\beta^* t(1-t)}dt} & : \quad \mu_1^* = 0.5, \mu_2^* \neq \frac{1}{3} \\ \frac{\beta^* e^{\left(\frac{\beta^*}{b-a}\right)y}}{(b-a)\left(e^{\beta^*}-1\right)e^{\frac{a\beta^*}{b-a}}} & : \mu_1^*, \mu_2^* \text{ for monotonicity} \\ \frac{e^{-\frac{1}{2}\left(\frac{y-\mu_Y^{(1)}}{\sigma_Y^{(1)}}\right)^2}}{(b-a)\int_0^1 e^{-\frac{1}{2}\left(\frac{x-\mu_1^*}{\sigma}\right)^2}dx} & : \text{for a small variance} \\ \frac{1}{b-a}\frac{\exp\left(\beta^*\left(\frac{y-a}{b-a}\right)+\gamma^*\left(\frac{y-a}{b-a}\right)^2\right)}{\int\limits_0^1 e^{\beta^* t+\gamma^* t^2}dt} & : \quad \textbf{the non special case} \end{cases} \tag{12.137}$$

$$a \leq y \leq b$$

12.7.2 The input- and the solution space

As we have already discussed in the subsection of algorithmic steps, the first two moments of the desired minimum information probability distribution are given as inputs. Effectively, the first two moments as inputs necessary for the numerical solution of the system (12.130) are μ_1^* and μ_2^* respectively.

Only on finding the numerical solution of the system (12.130), our primary objective of determining the desired minimum information probability distribution will be served. On the other hand, any numerical solution to the given system of simultaneous equations necessitates the complete knowledge of

- the **input space**, which consists of all possible pairs (μ_1^*, μ_2^*), for which the aforesaid numerical solutions do exist
- the **solution space**, which consists of the possible solutions achievable as a result of inputs belonging to the input space

Since the uniqueness of the solution of (6.8) has already been established, the unique solution of (12.130) will be (β^*, γ^*) corresponding to the given input (μ_1^*, μ_2^*). This means, by the definitions of $\mu_1^{(\beta,\gamma)}$ (from 12.85) and $\mu_2^{(\beta,\gamma)}$ (from 12.86), we have $\mu_1^{(\beta^*,\gamma^*)} = \mu_1^*$ and $\mu_2^{(\beta^*,\gamma^*)} = \mu_2^*$. It is also known, that for every μ_1^*, we must have $0 < \mu_1^* < 1$ and for that,

- ${\mu_1^*}^2 < \mu_2^* < \mu_1^*$ and
- $-\infty < \beta^* < \infty \;\&\; -\infty < \gamma^* < \infty$

which means, that

- the input space symbolized by $R_{\mu_1^*,\mu_2^*}$, the geometrical figure of which is given in the section 5.10, is an open and bounded set, which is geometrically represented by a finite area on the following two dimensional $\mu_1 - \mu_2$ plane, such that μ_1 stands for μ_1^* on the horizontal axis and μ_2 stands for μ_2^* on the vertical axis.

  This finite area containing all the possible input points is fully contained in the square on the $\mu_1 - \mu_2$ plane described by the corner points (0,0), (1,0), (1,1) and (0,1)

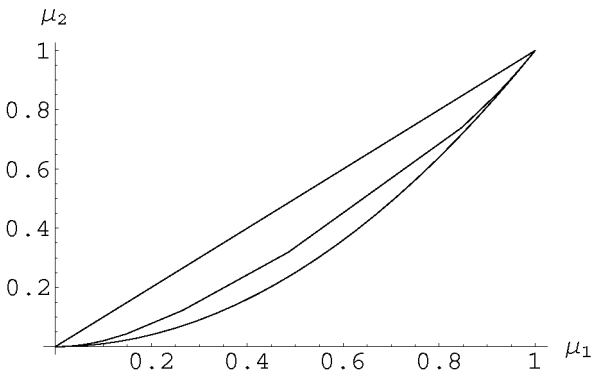


Figure 12.2: μ_1 against μ_2

- the solution space symbolized by $S_{(\beta^*,\gamma^*)}$ is an open and unbounded set, which can be geometrically represented by a two dimensional $\beta-\gamma$ plane, where β stands for β^* on the horizontal axis and γ stands for γ^* on the vertical axis, such that $(\beta^*, \gamma^*) \in \mathbb{R}^2$

  In the above figure, the curved line joining the points (0,0) and (1,1), which lies between the straight line segment "$\mu_2 = \mu_1$" and the parabolic arc "$\mu_2^2 = \mu_1$", represents the relationship between μ_1 and μ_2 in cases, where $\gamma = 0$

Our imminent discussions will be based on the fact, that every user-given input (μ_1^*, μ_2^*) in the input space has an unique solution (β^*, γ^*) as it's image in the solution space.

12.7.3 Characteristics of the outputs with respect to the inputs

The inputs are basically the first two user-given moments, namely μ_1^* and μ_2^*, whereas the outputs meant for the determination of the desired probability distribution are basically β^* and γ^*.

The solution (β^*, γ^*) of the system (12.130) can only be achieved by choosing a suitable first approximated solution (β_0, γ_0) of (12.130) meant for the Newton Raphson numerical procedure at first. The choice of a suitable (β_0, γ_0)

becomes extremely decisive in most of the cases. Though the Newton Raphson procedure is extremely robust, a wrong choice of (β_0, γ_0) may lead to a hopeless malfunctioning of this numerical procedure.

Before we go for the procedures for finding a (β_0, γ_0) and judge it's suitability, we shall have a look at certain important characteristics of the system (12.130). For this, we consider a two dimensional $\gamma - \mu_2$ curve, where $\mu_2 = \mu_2^{(\beta,\gamma)}$ (on the vertical axis) is plotted against γ (on the horizontal axis) for a fixed value of $\mu_1^{(\beta,\gamma)} = \mu_1^*$. The fixed value of μ_1^* $(0 < \mu_1^* < 1)$ determines the value of β for a given value of γ uniquely.

A $\gamma - \mu_2$ curve for any fixed value of $\mu_1^{(\beta,\gamma)} = \mu_1^*$ clearly shows, that there are two points of inflection on it. For the sake of clarification of the same, let us take an example of the $\gamma - \mu_2$ curve for $\mu_1^* = 0.525$

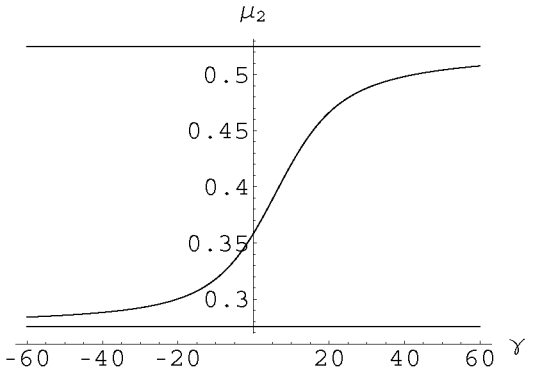


Figure 12.3: γ against μ_2

The horizontal lines $\mu_2 = 0.525$ and $\mu_2 = 0.525^2 = 0.275625$ in the above figure are the asymptotes of the curve.

The figure clearly shows, that the two points of inflection, namely one for $\gamma > 0$ called the upper point of inflection and the other for $\gamma < 0$ called the lower point of inflection, play a predominant role in the determination of the rate of change of μ_2 with respect to γ in different sections of the curve. There are three such sections, which we shall discuss right now.

Generally, for every μ_1^* $(0 < \mu_1^* < 1)$, if the upper and the lower points of inflection are symbolized by $(\gamma_{low}, \mu_{2low})$ and $(\gamma_{upp}, \mu_{2upp})$ respectively, then

we arrive at the following characteristics

- within the range of $\mu_2 > \mu_{2upp}$, γ tends to ∞ rapidly with the increase in μ_2. In fact, $\gamma \to \infty$ for $\mu_2 \to \mu_1^*$
- within the range of $\mu_2 < \mu_{2low}$, γ tends to $-\infty$ rapidly with the decrease in μ_2. In fact,$\gamma \to -\infty$ for $\mu_2 \to \mu_1^{*2}$
- within the range of $\mu_{2low} \leq \mu_2 \leq \mu_{2upp}$, γ changes slowly

Analytically, for a given μ_1^*, γ_{upp} and γ_{low} are the roots of the following equation

$$\frac{d^2}{d\gamma^2}\left\{\mu_2^{(\beta,\gamma)}\right\} = 0 \tag{12.138}$$

such that β is a function of γ and is connected by $\mu_1^{(\beta,\gamma)} = \mu_1^*$ implicitly.

It can be easily seen, that the computations of γ_{upp} and γ_{low} require numerical differentiation, which involves a good amount of numerical work. Only if the following sufficient conditions

$$\frac{d^3}{d\gamma^3}\left\{\mu_2^{(\beta,\gamma)}\right\}\Big|_{\beta=\beta_{upp},\gamma=\gamma_{upp}} \neq 0 \tag{12.139}$$

$$\frac{d^3}{d\gamma^3}\left\{\mu_2^{(\beta,\gamma)}\right\}\Big|_{\beta=\beta_{low},\gamma=\gamma_{low}} \neq 0 \tag{12.140}$$

hold such that

$$\mu_1^{(\beta_{upp},\gamma_{upp})} = \mu_1^{(\beta_{low},\gamma_{low})} = \mu_1^* \tag{12.141}$$

then we can compute

$$\mu_{2upp} = \mu_2^{(\beta_{upp},\gamma_{upp})} \ \& \ \mu_{2low} = \mu_2^{(\beta_{low},\gamma_{low})} \tag{12.142}$$

For our work, the exact numerical computations of $(\gamma_{low}, \mu_{2low})$ and $(\gamma_{upp}, \mu_{2upp})$ are not really necessary, though rough estimations of the same would always be of a big guidance for the programmers, as far as the choice of a suitable (β_0, γ_0) for a given (μ_1^*, μ_2^*) is concerned.

This is to say, that, if $\mu_2^* > \mu_{2upp}$ or $\mu_2^* < \mu_{2low}$, then the choice of (β_0, γ_0) needs to be handled carefully, otherwise the Newton Raphson procedure cannot be successfully executed. This is simply because, within these ranges of μ_2^*, the variability range of γ_0 is expectedly big.

However, in certain cases of $\mu_2^* < \mu_{2low}$, if the the variance $\sigma^2 = \mu_2^* - \mu_1^{*2}$ fulfills the following condition

$$\sigma < \frac{\mu_D}{\ell}, \quad \mu_D = \min\{\mu_1^*, 1 - \mu_1^*\} \tag{12.143}$$

which means

$$\mu_2^* < \mu_1^{*2}\left(1 + \frac{1}{\ell^2}\right), \text{ if } \mu_1^* \leq 0.5 \tag{12.144}$$

$$\mu_2^* < \mu_1^{*2}\left(1 + \frac{1}{\ell^2}\right) + \frac{1 - 2\mu_1^*}{\ell^2}, \text{ if } \mu_1^* > 0.5 \tag{12.145}$$

then no choice of (β_0, γ_0) is necessary anymore, since an approximating truncated normal probability distribution for a small variance (as already discussed in one of the preceding subsections) will be the final result for the user.

But, if $\mu_{2low} \leq \mu_2^* \leq \mu_{2upp}$, then the choice of a suitable (β_0, γ_0) is relatively simple and the Newton Raphson procedure takes only seconds to compute the final (β^*, γ^*).

Lastly, it must be unforgettably stated, that for any fixed value of $\mu_1^{(\beta,\gamma)} = \mu_1^*$, the following holds: $\beta \to \pm\infty \Leftrightarrow \gamma \to \mp\infty$. This fact has already been established by $\frac{d\beta}{d\gamma} < 0$ for a fixed μ_1 in (6.25).

12.7.4 An useful transformation $Z = 1 - X$

My programming trials have shown, that in certain sectional areas of $R_{\mu_1^*,\mu_2^*}$, the user-given input (μ_1^*, μ_2^*) cannot be processed or processed very easily by the program to give the final (β^*, γ^*). This very problem mainly exists, when the variance $\sigma^2 = \mu_2^* - \mu_1^{*2}$ is small but not small enough for truncated normal approximations in the areas where μ_1^* exceeds 0.875. Additionally, even if $\mu_1 \in (0.5, 0.875)$, then the computing speed as a result of this useful transformation is considerably increased in certain areas.

It has to be clearly noted, that the existence of this problem must also be adjudged on the basis of the running speed of the program corresponding to the particular user-given input (μ_1^*, μ_2^*). With reference to the experienced computing speeds of the program together with the knowledge of the other inevitable problems, these sectional areas are enlisted as follows:

- $\{(\mu_1^*, \mu_2^*) \mid \mu_1^* > 0.99\}$
- $\{(\mu_1^*, \mu_2^*) \mid 0.5 < \mu_1^* < 0.875 \ , \sigma^2 < 0.04\}$
- $\{(\mu_1^*, \mu_2^*) \mid 0.875 \leq \mu_1^* \leq 0.99 \ , \sigma^2 < 0.00128\}$
- $\{(\mu_1^*, \mu_2^*) \mid 0.95 < \mu_1^* < 0.975 \ , 0.9025 < \mu_2^* < 0.9628125\}$
- $\{(\mu_1^*, \mu_2^*) \mid 0.975 < \mu_1^* < 0.99 \ , 0.950625 < \mu_2^* < 0.98505\}$

(12.146)

In such cases, if the random variable X is transformed to the random variable $Z = 1 - X$, such that

- $E[X] = \mu_1^*$ and
- $E[X^2] = \mu_2^*$ & $Var[X] = \mu_2^* - \mu_1^{*2}$,

we must have

- $E[Z] = 1 - \mu_1^*$ and
- $E[Z^2] = 1 - 2\mu_1^* + \mu_2^*$ & $Var[Z] = \mu_2^* - \mu_1^{*2}$

The purpose of this transformation is to feed the program with the input $d_Z = (1 - \mu_1^*, 1 - 2\mu_1^* + \mu_2^*)$ instead of the input $d = (\mu_1^*, \mu_2^*)$. That is, the program uses the first two moments of the random variable Z instead of the same of the random variable X, but achieve exactly the desired output (β^*, γ^*) finally.

Experimentally it has been found, that this transformation is not necessary or rather harmful in the rest of the areas of $R_{\mu_1^*, \mu_2^*}$.

Now, our objective in this subsection will be to derive the theory of this transformation, with the help of which our program shall be developed for the aforesaid affected subareas of $R_{\mu_1^*, \mu_2^*}$.

The probability density functions of X and Z, which are $f_{X|\{d\}}(x)$, $0 \leq x \leq 1$ and $f_{Z|\{d_Z\}}(z)$, $0 \leq z \leq 1$ respectively, are given as

$$f_{X|\{d\}}(x) = B^* e^{\beta^* x + \gamma^* x^2}, \quad 0 \leq x \leq 1 \tag{12.147}$$

$$f_{Z|\{d_Z\}}(z) = \overline{B^*} e^{\overline{\beta^*} z + \overline{\gamma^*} z^2}, \quad 0 \leq z \leq 1 \tag{12.148}$$

which leads us to the following

$$B^* = \frac{1}{\int\limits_0^1 e^{\beta^* x+\gamma^* x^2}dx} \;\&\; \overline{B^*} = \frac{1}{\int\limits_0^1 e^{\overline{\beta^*} x+\overline{\gamma^*} x^2}dx} \tag{12.149}$$

$$f_{Z|\{d_Z\}}(z) = f_{X|\{d\}}(x)\left|\frac{dx}{dz}\right| = f_{X|\{d\}}(x) \tag{12.150}$$

Therefore, the program, which accepts $(1-\mu_1^*, 1-2\mu_1^*+\mu_2^*)$ as the input, gives $(\overline{\beta^*}, \overline{\gamma^*})$ as the output. In other words, the program solves the following equation at first:

$$\begin{cases} \dfrac{\int\limits_0^1 xe^{\beta x+\gamma x^2}dx}{\int\limits_0^1 e^{\beta x+\gamma x^2}dx} = 1-\mu_1^* \\ \dfrac{\int\limits_0^1 x^2e^{\beta x+\gamma x^2}dx}{\int\limits_0^1 e^{\beta x+\gamma x^2}dx} = 1-2\mu_1^*+\mu_2^* \end{cases} \tag{12.151}$$

and gives $(\overline{\beta^*}, \overline{\gamma^*})$ as the solution of the same. Only after this, (β^*, γ^*), which is the solution of (12.130), is given as the final output by the procedure of the reverse transformation described by $X = 1 - Z$.

Hence, in order to find the theoretical rule for this reverse transformation, our aim will be to derive the expressions of β^*, γ^* and B^* in terms of the known values of $\overline{\beta^*}$, $\overline{\gamma^*}$ and $\overline{B^*}$ yielded by the program.

By using (12.150), we get

$$\begin{aligned} \overline{B^*}e^{\overline{\beta^*}z+\overline{\gamma^*}z^2} &= B^*e^{\beta^* x+\gamma^* x^2} = B^*e^{\beta^*(1-z)+\gamma^*(1-z)^2} \\ &= B^*e^{\beta^*+\gamma^*-(\beta^*+2\gamma^*)z+\gamma^* z^2} \\ &= B^*e^{\beta^*+\gamma^*}e^{-(\beta^*+2\gamma^*)z+\gamma^* z^2} \end{aligned} \tag{12.152}$$

and therefore by comparing the coefficients of z on both the sides of the above relation, we get

$$\begin{aligned} \overline{B^*} &= B^*e^{\beta^*+\gamma^*} \\ \overline{\beta^*} &= -(\beta^*+2\gamma^*) \\ \overline{\gamma^*} &= \gamma^* \end{aligned} \tag{12.153}$$

which gives in turn

$$\begin{aligned} B^* &= \overline{B^*} e^{\overline{\beta^*}+\overline{\gamma^*}} \\ \beta^* &= -(\overline{\beta^*} + 2\overline{\gamma^*}) \\ \gamma^* &= \overline{\gamma^*} \end{aligned} \tag{12.154}$$

Formally, in order to cover the possibilities of overflow errors, B^* is computed by the following rule

$$B^* = \begin{cases} \overline{B^*} e^{\overline{\beta^*}+\overline{\gamma^*}} & : \quad \overline{\beta^*} + \overline{\gamma^*} < 709 \\ \frac{1}{\int\limits_0^1 e^{\overline{\beta^*}(x-1)+\overline{\gamma^*}(x^2-1)} dx} & : \quad \overline{\beta^*} + \overline{\gamma^*} \geq 709 \end{cases}$$

Whence, (12.154) gives us our desired solution (β^*, γ^*) and subsequently the desired probability distribution $f_{X|\{d\}}(x)$, $0 \leq x \leq 1$.

12.7.5 The data structure

A data access is absolutely necessary for the computation of the start vector (β_0, γ_0) for an user-given input (μ_1^*, μ_2^*) in cases, when $0.01 \leq \mu_1^* \leq 0.99$ and the condition (11.29) is not fulfilled. This means, if

$$(\mu_1^*, \mu_2^*) \in \left\{ (\mu_1, \mu_2) \mid 0.01 \leq \mu_1 \leq 0.99\,, \mu_2 \geq \mu_1^2 + \frac{\min\{\mu_1, 1-\mu_1\}}{\ell} \right\} \tag{12.155}$$

then the subroutine of the program, which computes the solution (β^*, γ^*) in these cases, needs to access a programmer generated data.

It must be unforgettably stated, that ℓ does not have a fixed value, but varies with different ranges of μ_1^*. These different values of ℓ are carefully chosen to make sure, that the best possible results could be yielded. These values of ℓ shall be given in due course.

Our database[2] consists of about twenty-six thousand records. Each record contains four fields (*columns*), namely γ, β, $\mu_1^{(\beta,\gamma)}$ and $\mu_2^{(\beta,\gamma)}$, the symbols of which have their usual meanings. In this very context, accessing the database means, that the subroutine accesses records ,with subject to certain programmer specified access rules. These records (four, two or one at a time, as the case maybe) shall be accessed by our software program.

Corresponding to certain specified values of $\mu_1^{(\beta,\gamma)} = \mu_1$, i.e.

$$\begin{aligned} \mu_1 \in mep_{category} = \{&0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05, \\ &0.055, 0.06, 0.065, 0.07, 0.075, 0.08, 0.085, 0.09, 0.095, \\ &0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25, 0.275, 0.3, \\ &0.325, 0.35, 0.375, 0.4, 0.425, 0.45, 0.475, 0.5 \\ &0.525, 0.55, 0.575, 0.6, 0.625, 0.65, 0.675, 0.7 \\ &0.725, 0.75, 0.775, 0.8, 0.825, 0.85, 0.875, 0.9, \\ &0.925, 0.95, 0.975, 0.99\} \end{aligned}$$

the records contained in the entire database are divided into several categories in a way, that each category C_{μ_1} is keyed by μ_1, ie. one of the values of

[2]In our context, database means the complete set of files containing all the start vectors. This complete set shall be termed as the complete data or precisely database by us.

$mep_{category}$. As for example, the category $C_{0.6}$ is a set of the records given by the following description:

$$C_{0.6} = \left\{ \left(\gamma, \beta, 0.6, \mu_2^{(\beta,\gamma)}\right) \left| \frac{\int\limits_0^1 x e^{\beta x+\gamma x^2} dx}{\int\limits_0^1 e^{\beta x+\gamma x^2} dx} = 0.6 \quad \& \quad \frac{\int\limits_0^1 x^2 e^{\beta x+\gamma x^2} dx}{\int\limits_0^1 e^{\beta x+\gamma x^2} dx} = \mu_2^{(\beta,\gamma)} \right. \right\}$$

It can be easily noted, that there are 55 categories altogether.

Therefore, we symbolize

$$mep2 = \bigcup_{\mu_1 \in mep_{category}} C_{\mu_1}$$

In the subsequent subsection, we shall discuss about the procedures, by which $mep2$ should be accessed (i.e. the procedures for the database access). An $mep2$ access gives an access vector (β_a, γ_a), the further processing of which yields the starting vector (β_0, γ_0) meant for the Newton Raphson procedure.

12.7.6 Database access procedures for start vectors

Under the assumption, that the database access is necessary corresponding to the user-given input (μ_1^*, μ_2^*) with subject to the fulfillment of the condition (12.155), the $mep2$ access procedure (which shall be eventually categorized into certain subcases) has been developed with the reference to the picture of the bounded input space (figure 12.2).

Within the input space, the user-given (μ_1^*, μ_2^*) fulfilling (12.155) is enclosed by a smallest possible quadrilateral $PQRS$ in a way, that the coordinates of the points P, Q, R and S are the real numbers saved in the third and the fourth field of some of the chosen available records belonging to $mep2$.

The quadrilateral $\Box PQRS$ is a trapezium, whose the segments $\overline{SP}$ and $\overline{QR}$ parallel to each other. Obviously, the enveloping trapezium $\Box PQRS$ is completely contained in the input space, the picture of whose is given in the figure 12.4.

It is important to note, that (μ_1^*, μ_2^*) may be an interior point or a boundary point (may lie on the segment $\overline{SP}$) of $\Box PQRS$.

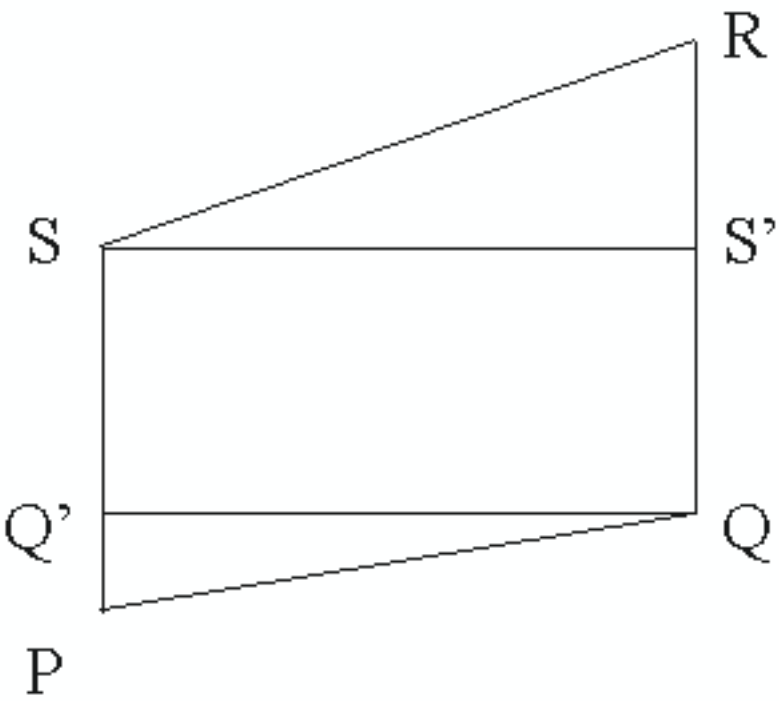


Figure 12.4: Enveloping Trapezium $\Box PQRS$

Now, our immediate problem will be to find out a meaningful procedure of choosing our $\Box PQRS$, with subject to the user-given input (μ_1^*, μ_2^*), so that an useful access vector (β_a, γ_a) can be yielded eventually.

At the very first step, corresponding to the given μ_1^*, two values of the first moment μ_1, namely $\mu_1^{(Q)}$, $\mu_1^{(P)} \in mep_{category}$ are selected in a way, that

- $\mu_1^{(Q)} = \inf_{\mu_1 \geq \mu_1^*} \mu_1$
- $\mu_1^{(P)} = \sup_{\mu_1 \leq \mu_1^*} \mu_1$

Here, two possible cases do arise, namely $\mu_1^{(Q)} = \mu_1^{(P)}$ or $\mu_1^{(Q)} > \mu_1^{(P)}$:

1. $\mu_1^{(Q)} = \mu_1^{(P)}$

   This means, $\mu_1^{(Q)} = \mu_1^{(P)} = \mu_1^* \in mep_{category}$. In the next step of this case, select two records $(\gamma^{(S)}, \beta^{(S)}, \mu_1^*, \mu_2^{(S)})$ and $(\gamma^{(P)}, \beta^{(P)}, \mu_1^*, \mu_2^{(P)})$ from $C_{\mu_1^*}$ in a way, that

   - $\mu_2^{(S)} = \inf_{\mu_2 \geq \mu_2^*} \mu_2$
   - $\mu_2^{(P)} = \sup_{\mu_2 \leq \mu_2^*} \mu_2$

   Again, this very case must be necessarily subdivided into two subcases, namely $\mu_2^{(S)} = \mu_2^{(P)}$ or $\mu_2^{(S)} > \mu_2^{(P)}$:

   (a) $\mu_2^{(S)} = \mu_2^{(P)}$: In this case, the records $(\gamma^{(P)}, \beta^{(P)}, \mu_1^*, \mu_2^{(P)})$ and $(\gamma^{(S)}, \beta^{(S)}, \mu_1^*, \mu_2^{(S)})$ are identical.

   Therefore, we simply take $\beta_a = \beta^{(P)}$ and $\gamma_a = \gamma^{(P)}$ and thereby our access vector will be (β_a, γ_a).

   This is the case, where only **one** record is needed to be accessed for choosing our access vector, where $\square PQRS$ turns out to be a single point P

   (b) $\mu_2^{(S)} > \mu_2^{(P)}$: In this case, the records $(\gamma^{(P)}, \beta^{(P)}, \mu_1^*, \mu_2^{(P)})$ and $(\gamma^{(S)}, \beta^{(S)}, \mu_1^*, \mu_2^{(S)})$ are distinct.

   Therefore, our access vector will be yielded according to the following rule:

   - if $(\mu_2^{(S)} - \mu_2^*) \geq (\mu_2^* - \mu_2^{(P)})$ then $\beta_a = \beta^{(P)}$ and $\gamma_a = \gamma^{(P)}$
   - if $(\mu_2^{(S)} - \mu_2^*) < (\mu_2^* - \mu_2^{(P)})$ then $\beta_a = \beta^{(S)}$ and $\gamma_a = \gamma^{(S)}$

   This is the case, where only **two** records are needed to be accessed for choosing our access vector, where $\square PQRS$ turns out to be a segment $\overline{SP}$ parallel to the vertical axis

2. $\mu_1^{(Q)} > \mu_1^{(P)}$

   This means, $\mu_1^* \notin mep_{category}$. In the next step of this case, since it is clearly known that

   - $\mu_1^{(left)} = \mu_1^{(P)} = \mu_1^{(S)}$ and
   - $\mu_1^{(right)} = \mu_1^{(Q)} = \mu_1^{(R)}$,

   the two categories $C_{\mu_1^{(P)}}$ and $C_{\mu_1^{(Q)}}$ can be presented in the following way

   - $C^{(left)} = C_{\mu_1^{(P)}} = C_{\mu_1^{(S)}}$
   - $C^{(right)} = C_{\mu_1^{(Q)}} = C_{\mu_1^{(R)}}$

   after which two records $(\gamma^{(left)}, \beta^{(left)}, \mu_1^{(S)}, \mu_2^{(left,\max)}) \in C^{(left)}$ and $(\gamma^{(right)}, \beta^{(right)}, \mu_1^{(Q)}, \mu_2^{(right,\min)}) \in C^{(right)}$ are selected in the following way

   - $\mu_2^{(left,\max)} = \sup_{\mu_1 \in C^{(left)}} \mu_2$ and is slightly smaller than $\mu_1^{(left)}$
   - $\mu_2^{(right,\min)} = \inf_{\mu_1 \in C^{(right)}} \mu_2$ and is slightly greater than $\mu_1^{(right)^2}$

   The geometrical view of the input space (figure 12.2) clearly tells us, that for every chosen (μ_1^*, μ_2^*),

   - $\mu_2^{(left,\max)} < \sup_{\mu_1 \in C^{(right)}} \mu_2$ and
   - $\mu_2^{(right,\min)} > \inf_{\mu_1 \in C^{(left)}} \mu_2$

   This explains, why we have constructed our $\Box PQRS$ (figure 12.4) in a way, that

   - ordinate of S = ordinate of $S^{'}$ < ordinate of R
   - ordinate of Q = ordinate of $Q^{'}$ > ordinate of P

   For our further course of discussions, the introduction of a vectorial rule, named *counter clockwise rule*, is vitally important.

This rule determines the position of a point (on a plane) with respect to a given directed line. This directed line is determined by two points belonging to it. By position of the point we mean, whether the point is on the left or on the right side of the directed line or on the line itself:

Assuming that the points A and B have the coordinates (a_1, a_2) and (b_1, b_2) respectively, then the point $D : (d_1, d_2)$ is said to be on the left or the right side of the directed line containing $\overrightarrow{AB}$ according as

$$CCW(A, B, D) = \begin{vmatrix} a_1 & a_2 & 1 \\ b_1 & b_2 & 1 \\ d_1 & d_2 & 1 \end{vmatrix} \gtrless 0 \tag{12.156}$$

Additionally, D is said to lie on the directed line containing $\overrightarrow{AB}$, if

$$CCW(A, B, D) = 0$$

At this point, the coordinates of P, Q, R and S can be chosen to be $(\mu_1^{(left)}, \mu_2^{(P)})$, $(\mu_1^{(right)}, \mu_2^{(Q)})$, $(\mu_1^{(right)}, \mu_2^{(R)})$ and $(\mu_1^{(left)}, \mu_2^{(S)})$ respectively, with the help of which a suitable construction of $\Box PQRS$ would be made possible by means of a proper choice of four records, namely $(\gamma^{(P)}, \beta^{(P)}, \mu_1^{(left)}, \mu_2^{(P)})$, $(\gamma^{(Q)}, \beta^{(Q)}, \mu_1^{(right)}, \mu_2^{(Q)})$, $(\gamma^{(R)}, \beta^{(R)}, \mu_1^{(right)}, \mu_2^{(R)})$ and $(\gamma^{(S)}, \beta^{(S)}, \mu_1^{(left)}, \mu_2^{(S)})$ by database access.

The development of this database access algorithmic rule necessitates the subdivision of the case $(\mu_1^{(Q)} > \mu_1^{(P)})$ into three subcases, namely

- $\mu_2^* \leq \mu_2^{(right,\min)}$
- $\mu_2^* \geq \mu_2^{(left,\max)}$
- $\mu_2^{(right,\min)} < \mu_2^* < \mu_2^{(left,\max)}$

Before we discuss the above subcases individually, we must keep in mind, that the quantities, namely $\mu_1^{(left)}$, $\mu_1^{(right)}$, $\mu_2^{(left,\max)}$ and $\mu_2^{(right,\min)}$ are already known.

Since the selection of the abscissae, namely $\mu_1^{(left)}$ and $\mu_1^{(right)}$ of the points P, Q, R and S has already taken place, only the ordinates of the same are needed to be suitably chosen. The procedures for the selection of these ordinates are different in different subcases, which are discussed step-by-step as follows:

(a) $\mu_2^* \le \mu_2^{(right,\min)}$

- Select the record $(\gamma^{(Q)}, \beta^{(Q)}, \mu_1^{(right)}, \mu_2^{(Q)}) \in C^{(right)}$ such that
$$\mu_2^{(Q)} = \mu_2^{(right,\min)}$$
The selected record is therefore $(\gamma^{(Q)}, \beta^{(Q)}, \mu_1^{(right)}, \mu_2^{(right,\min)})$, the coordinates of the point Q being $(\mu_1^{(right)}, \mu_2^{(right,\min)})$
- At first, the selection of a record $(\gamma^{(P)}, \beta^{(P)}, \mu_1^{(left)}, \mu_2^{(P)}) \in C^{(left)}$ must be subjected to the fact, that the point (μ_1^*, μ_2^*) must lie on the left of the ray $\overrightarrow{PQ}$. This means, by the rule (12.156),
$$det_{\overrightarrow{PQ}} = \begin{vmatrix} \mu_1^{(left)} & \mu_2^{(P)} & 1 \\ \mu_1^{(right)} & \mu_2^{(Q)} & 1 \\ \mu_1^* & \mu_2^* & 1 \end{vmatrix} > 0 \qquad (12.157)$$
$$\Leftrightarrow \begin{vmatrix} \mu_1^{(left)} - \mu_1^* & \mu_1^{(P)} - \mu_2^* \\ \mu_1^{(right)} - \mu_1^* & \mu_1^{(Q)} - \mu_2^* \end{vmatrix} > 0$$
$$\Leftrightarrow (\mu_1^{(left)} - \mu_1^*)(\mu_2^{(Q)} - \mu_2^*) > (\mu_1^{(right)} - \mu_1^*)(\mu_2^{(P)} - \mu_2^*)$$
$$\Leftrightarrow \mu_2^{(P)} < \mu_2^* + \frac{(\mu_1^{(left)} - \mu_1^*)(\mu_2^{(Q)} - \mu_2^*)}{\mu_1^{(right)} - \mu_1^*} = \overline{\mu}_2 \qquad (12.158)$$
with subject to the fact, that $\mu_1^{(right)} > \mu_1^*$.

  The selected record is therefore $(\gamma^{(P)}, \beta^{(P)}, \mu_1^{(left)}, \mu_2^{(P)})$, such that
$$\mu_2^{(P)} = \sup_{\mu_2 < \overline{\mu}_2} \mu_2 = \sup_{det_{\overrightarrow{PQ}} > 0} \mu_2,$$
  the coordinates of the point P being $(\mu_1^{(left)}, \sup\limits_{det_{\overrightarrow{PQ}} > 0} \mu_2)$
- Select the record $(\gamma^{(R)}, \beta^{(R)}, \mu_1^{(right)}, \mu_2^{(R)}) \in C^{(right)}$ such that
$$\mu_2^{(R)} = \inf_{\mu_2 > \mu_2^{(Q)}} \mu_2$$
  The selected record is therefore $(\gamma^{(R)}, \beta^{(R)}, \mu_1^{(right)}, \inf\limits_{\mu_2 > \mu_2^{(Q)}} \mu_2)$, the coordinates of the point R being $(\mu_1^{(right)}, \inf\limits_{\mu_2 > \mu_2^{(Q)}} \mu_2)$

- Select the record $(\gamma^{(S)}, \beta^{(S)}, \mu_1^{(left)}, \mu_2^{(S)}) \in C^{(left)}$ such that

$$\mu_2^{(S)} = \inf_{\mu_2 > \mu_2^*} \mu_2$$

The selected record is therefore $(\gamma^{(S)}, \beta^{(S)}, \mu_1^{(left)}, \inf_{\mu_2 > \mu_2^*} \mu_2)$, the coordinates of the point S being $(\mu_1^{(left)}, \inf_{\mu_2 > \mu_2^*} \mu_2)$

This completes the construction of $\square PQRS$ in case of $\mu_2^* \leq \mu_2^{(right,\min)}$

(b) $\mu_2^* \geq \mu_2^{(left,\max)}$

- Select the record $(\gamma^{(S)}, \beta^{(S)}, \mu_1^{(left)}, \mu_2^{(S)}) \in C^{(left)}$ such that

$$\mu_2^{(S)} = \mu_2^{(left,\max)}$$

The selected record is therefore $(\gamma^{(S)}, \beta^{(S)}, \mu_1^{(left)}, \mu_2^{(left,\max)})$, the coordinates of the point S being $(\mu_1^{(left)}, \mu_2^{(left,\max)})$

- At first, the selection of a record $(\gamma^{(R)}, \beta^{(R)}, \mu_1^{(right)}, \mu_2^{(R)}) \in C^{(right)}$ must be subjected to the fact, that the point (μ_1^*, μ_2^*) must lie on the right of the ray $\overrightarrow{SR}$. This means, by the rule (12.156),

$$det_{\overrightarrow{SR}} = \begin{vmatrix} \mu_1^{(left)} & \mu_2^{(S)} & 1 \\ \mu_1^{(right)} & \mu_2^{(R)} & 1 \\ \mu_1^* & \mu_2^* & 1 \end{vmatrix} < 0 \tag{12.159}$$

$$\Leftrightarrow \quad \begin{vmatrix} \mu_1^{(left)} - \mu_1^* & \mu_1^{(S)} - \mu_2^* \\ \mu_1^{(right)} - \mu_1^* & \mu_1^{(R)} - \mu_2^* \end{vmatrix} < 0$$

$$\Leftrightarrow \quad (\mu_1^{(left)} - \mu_1^*)(\mu_2^{(R)} - \mu_2^*) < (\mu_1^{(right)} - \mu_1^*)(\mu_2^{(S)} - \mu_2^*)$$

$$\Leftrightarrow \quad \mu_2^{(R)} > \mu_2^* + \frac{(\mu_1^{(right)} - \mu_1^*)(\mu_2^{(S)} - \mu_2^*)}{\mu_1^{(left)} - \mu_1^*} = \overline{\mu}_2 \tag{12.160}$$

with subject to the fact, that $\mu_1^{(left)} < \mu_1^*$.

The selected record is therefore $(\gamma^{(R)}, \beta^{(R)}, \mu_1^{(right)}, \mu_2^{(R)})$, such that

$$\mu_2^{(R)} = \inf_{\mu_2 > \overline{\mu}_2} \mu_2 = \inf_{det_{\overrightarrow{SR}} < 0} \mu_2,$$

the coordinates of the point R being $(\mu_1^{(right)}, \inf_{det_{\overrightarrow{SR}} < 0} \mu_2)$

- Select the record $(\gamma^{(P)}, \beta^{(P)}, \mu_1^{(left)}, \mu_2^{(P)}) \in C^{(left)}$ such that

$$\mu_2^{(P)} = \sup_{\mu_2 < \mu_2^{(S)}} \mu_2$$

The selected record is therefore $(\gamma^{(P)}, \beta^{(P)}, \mu_1^{(left)}, \sup_{\mu_2 < \mu_2^{(S)}} \mu_2)$, the coordinates of the point P being $(\mu_1^{(left)}, \sup_{\mu_2 < \mu_2^{(S)}} \mu_2)$

- Select the record $(\gamma^{(Q)}, \beta^{(Q)}, \mu_1^{(right)}, \mu_2^{(Q)}) \in C^{(right)}$ such that

$$\mu_2^{(Q)} = \sup_{\mu_2 < \mu_2^*} \mu_2$$

The selected record is therefore $(\gamma^{(Q)}, \beta^{(Q)}, \mu_1^{(right)}, \sup_{\mu_2 < \mu_2^*} \mu_2)$, the coordinates of the point Q being $(\mu_1^{(right)}, \sup_{\mu_2 < \mu_2^*} \mu_2)$

This completes the construction of $\Box PQRS$ in case of $\mu_2^* \geq \mu_2^{(left,\max)}$

(c) $\mu_2^{(right,\min)} < \mu_2^* < \mu_2^{(left,\max)}$

- Select the record $(\gamma^{(R)}, \beta^{(R)}, \mu_1^{(right)}, \mu_2^{(R)}) \in C^{(right)}$ such that

$$\mu_2^{(R)} = \inf_{\mu_2 \geq \mu_2^*} \mu_2$$

The selected record is therefore $(\gamma^{(R)}, \beta^{(R)}, \mu_1^{(right)}, \inf_{\mu_2 \geq \mu_2^*} \mu_2)$, the coordinates of the point R being $(\mu_1^{(right)}, \inf_{\mu_2 \geq \mu_2^*} \mu_2)$

- Select the record $(\gamma^{(Q)}, \beta^{(Q)}, \mu_1^{(right)}, \mu_2^{(Q)}) \in C^{(right)}$ such that

$$\mu_2^{(Q)} = \sup_{\mu_2 \leq \mu_2^*} \mu_2$$

The selected record is therefore $(\gamma^{(Q)}, \beta^{(Q)}, \mu_1^{(right)}, \sup_{\mu_2 \leq \mu_2^*} \mu_2)$, the coordinates of the point Q being $(\mu_1^{(right)}, \sup_{\mu_2 \leq \mu_2^*} \mu_2)$

- Select the record $(\gamma^{(S)}, \beta^{(S)}, \mu_1^{(left)}, \mu_2^{(S)}) \in C^{(left)}$ such that

$$\mu_2^{(S)} = \inf_{\mu_2 \geq \mu_2^*} \mu_2$$

The selected record is therefore $(\gamma^{(S)}, \beta^{(S)}, \mu_1^{(left)}, \inf_{\mu_2 \geq \mu_2^*} \mu_2)$, the coordinates of the point S being $(\mu_1^{(left)}, \inf_{\mu_2 \geq \mu_2^*} \mu_2)$

- Select the record $(\gamma^{(P)}, \beta^{(P)}, \mu_1^{(left)}, \mu_2^{(P)}) \in C^{(left)}$ such that

$$\mu_2^{(P)} = \sup_{\mu_2 \leq \mu_2^*} \mu_2$$

The selected record is therefore $(\gamma^{(P)}, \beta^{(P)}, \mu_1^{(left)}, \sup_{\mu_2 \leq \mu_2^*} \mu_2)$, the coordinates of the point P being $(\mu_1^{(left)}, \sup_{\mu_2 \leq \mu_2^*} \mu_2)$

This completes the construction of $\square PQRS$ in case of $\mu_2^{(right,\min)} < \mu_2^* < \mu_2^{(left,\max)}$.

It has to be noted, that in this case, the ordinate of R may be smaller than or equal to the same of that of S (or $S^{'}$) i.e $\mu_2^{(R)} \leq \mu_2^{(S)}$ and the ordinate of P may be greater than or equal to the same of that of Q (or $Q^{'}$) i.e $\mu_2^{(P)} \geq \mu_2^{(Q)}$.

Therefore, the combination of the individual subroutines of the above three subcases constitute the subroutine of the database access rule for $\mu_1^{(right)} = \mu_1^{(Q)} > \mu_1^{(P)} = \mu_1^{(left)}$.

This completes the selection of the required trapezium $\square PQRS$ for $\mu_1^{(Q)} > \mu_1^{(P)}$. However, the determination of the access vector, which still remains pending, can therefore be performed in the next step.

Within the $\square PQRS$, by using the rule (12.156), the point (μ_1^*, μ_2^*) lies

- within the triangle ΔSRQ, but not on the segment $\overline{SQ}$
- exactly on the segment $\overline{SQ}$
- within the triangle ΔSPQ, but not on the segment $\overline{SQ}$

according as

$$det_{\overrightarrow{SQ}} = CCW(S,Q,T) = \begin{vmatrix} \mu_1^{(left)} & \mu_2^{(S)} & 1 \\ \mu_1^{(right)} & \mu_2^{(Q)} & 1 \\ \mu_1^* & \mu_2^* & 1 \end{vmatrix} \gtreqless 0 \tag{12.161}$$

where the point $T : (\mu_1^*, \mu_2^*)$, which is termed as the target vector in the input space, has the coordinates (μ_1^*, μ_2^*).

Again, the procedures for the determination of the access vector (β_a, γ_a) are different in the different cases, namely

$$det_{\overrightarrow{SQ}} \gtreqless 0$$

since the $\square PQRS$ is needed to be divided into two triangles, namely ΔSRQ and ΔSPQ.

These procedures are therefore discussed one by one as follows:

Case 1 ($(\mu_1^*, \mu_2^*) \in \Delta SRQ \setminus \overline{SQ}$):

The vector (μ_1^*, μ_2^*) is expressible as a convex combination of the vectors $S : (\mu_1^{(left)}, \mu_2^{(S)})$, $R : (\mu_1^{(right)}, \mu_2^{(R)})$ and $Q : (\mu_1^{(right)}, \mu_2^{(Q)})$ in a way, that

$$\begin{aligned} &(\mu_1^*, \mu_2^*) = a_1 \left(\mu_1^{(left)}, \mu_2^{(S)}\right) + a_2 \left(\mu_1^{(right)}, \mu_2^{(R)}\right) + a_3 \left(\mu_1^{(right)}, \mu_2^{(Q)}\right) \\ &\text{such that} \\ &a_1 + a_2 + a_3 = 1 \text{ and } a_1, a_2, a_3 \geq 0 \end{aligned} \tag{12.162}$$

These non-negative values a_1, a_2 and a_3 can be evaluated by solving the following system of linear equations by crammer's rule of determinants

$$\begin{aligned} \mu_1^{(left)} a_1 + \mu_1^{(right)} a_2 + \mu_1^{(right)} a_3 &= \mu_1^* \\ \mu_2^{(S)} a_1 + \mu_2^{(R)} a_2 + \mu_2^{(Q)} a_3 &= \mu_2^* \\ a_1 + a_2 + a_3 &= 1 \end{aligned} \tag{12.163}$$

Therefore, with subject to $\mu_2^{(R)} > \mu_2^{(Q)}$ and $\mu_2^{(right)} > \mu_2^{(left)}$, we get the

solution of (12.163) as follows:

$$\begin{aligned} a_1 &= \frac{\begin{vmatrix} \mu_1^{(right)} & \mu_2^{(Q)} & 1 \\ \mu_1^{(right)} & \mu_2^{(R)} & 1 \\ \mu_1^{*} & \mu_2^{*} & 1 \end{vmatrix}}{\left(\mu_2^{(R)} - \mu_2^{(Q)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} = \frac{CCW(Q,R,T)}{\left(\mu_2^{(R)} - \mu_2^{(Q)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} \\ &> 0, \text{ since } T \text{ is on the left of } \overrightarrow{QR} \end{aligned} \tag{12.164}$$

$$\begin{aligned} a_2 &= \frac{\begin{vmatrix} \mu_1^{(left)} & \mu_2^{(S)} & 1 \\ \mu_1^{(right)} & \mu_2^{(Q)} & 1 \\ \mu_1^{*} & \mu_2^{*} & 1 \end{vmatrix}}{\left(\mu_2^{(R)} - \mu_2^{(Q)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} = \frac{CCW(S,Q,T)}{\left(\mu_2^{(R)} - \mu_2^{(Q)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} \\ &> 0, \text{ since } T \text{ is on the left of } \overrightarrow{SQ} \end{aligned} \tag{12.165}$$

$$\begin{aligned} a_3 &= \frac{\begin{vmatrix} \mu_1^{(right)} & \mu_2^{(R)} & 1 \\ \mu_1^{(left)} & \mu_2^{(S)} & 1 \\ \mu_1^{*} & \mu_2^{*} & 1 \end{vmatrix}}{\left(\mu_2^{(R)} - \mu_2^{(Q)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} = \frac{CCW(R,S,T)}{\left(\mu_2^{(R)} - \mu_2^{(Q)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} \\ &\geq 0, \text{ since } T \text{ is on the left of } \overrightarrow{RS} \text{ or on } \overline{RS} \text{ itself} \end{aligned} \tag{12.166}$$

If T is found to lie exactly on $\overline{RS}$, it is clear, that $\overline{RS}$ must be parallel to the horizontal axis of the input plane, where $\mu_2^{(R)} = \mu_2^{(S)}$.

As a result of the solution of (12.163), by using the selected records $_{(\gamma^{(S)},\beta^{(S)},\mu_1^{(left)},\mu_2^{(S)})}$, $_{(\gamma^{(R)},\beta^{(R)},\mu_1^{(right)},\mu_2^{(R)})}$ and $_{(\gamma^{(Q)},\beta^{(Q)},\mu_1^{(right)},\mu_2^{(Q)})}$ the access vector is given as (β_a, γ_a), such that

$$\begin{aligned} \beta_a &= a_1\beta^{(S)} + a_2\beta^{(R)} + a_3\beta^{(Q)} \\ \gamma_a &= a_1\gamma^{(S)} + a_2\gamma^{(R)} + a_3\gamma^{(Q)} \end{aligned} \tag{12.167}$$

Case 2 ($(\mu_1^*, \mu_2^*) \in \overline{SQ}$):

The vector (μ_1^*, μ_2^*) is expressible as a linear combination of the vectors $S : (\mu_1^{(left)}, \mu_2^{(S)})$ and $Q : (\mu_1^{(right)}, \mu_2^{(Q)})$ in the following way

$$(\mu_1^*, \mu_2^*) = a_1 \left(\mu_1^{(left)}, \mu_2^{(S)}\right) + a_2 \left(\mu_1^{(right)}, \mu_2^{(Q)}\right) \tag{12.168}$$

which means

$$\begin{aligned} \mu_1^{(left)} a_1 + \mu_1^{(right)} a_2 &= \mu_1^* \\ \mu_2^{(S)} a_1 + \quad \mu_2^{(Q)} a_2 &= \mu_2^* \end{aligned} \tag{12.169}$$

Here, if $\mu_1^{(left)} \mu_2^{(Q)} = \mu_1^{(right)} \mu_2^{(S)}$, then the access vector is given as (β_a, γ_a), such that

$$\begin{aligned} \beta_a &= \left(\beta^{(S)} + \beta^{(Q)}\right)/2 \\ \gamma_a &= \left(\gamma^{(S)} + \gamma^{(Q)}\right)/2 \end{aligned} \tag{12.170}$$

Otherwise, by

$$\begin{aligned} a_1 &= \frac{\mu_1^* \mu_2^{(Q)} - \mu_1^{(right)} \mu_2^*}{\mu_1^{(left)} \mu_2^{(Q)} - \mu_1^{(right)} \mu_2^{(S)}} \\ a_2 &= \frac{\mu_1^{(left)} \mu_2^* - \mu_1^* \mu_2^{(S)}}{\mu_1^{(left)} \mu_2^{(Q)} - \mu_1^{(right)} \mu_2^{(S)}} \end{aligned} \tag{12.171}$$

the access vector is given as (β_a, γ_a), such that

$$\begin{aligned} \beta_a &= a_1 \beta^{(S)} + a_2 \beta^{(Q)} \\ \gamma_a &= a_1 \gamma^{(S)} + a_2 \gamma^{(Q)} \end{aligned} \tag{12.172}$$

Case 3 ($(\mu_1^*, \mu_2^*) \in \Delta SPQ \setminus \overline{SQ}$):

The vector (μ_1^*, μ_2^*) is expressible as a convex combination of the vectors $S : (\mu_1^{(left)}, \mu_2^{(S)})$, $P : (\mu_1^{(left)}, \mu_2^{(P)})$ and $Q : (\mu_1^{(right)}, \mu_2^{(Q)})$ in a way, that

$$\begin{aligned} &(\mu_1^*, \mu_2^*) = a_1 \left(\mu_1^{(left)}, \mu_2^{(S)}\right) + a_2 \left(\mu_1^{(left)}, \mu_2^{(P)}\right) + a_3 \left(\mu_1^{(right)}, \mu_2^{(Q)}\right) \\ &\text{such that} \\ &a_1 + a_2 + a_3 = 1 \text{ and } a_1, a_2, a_3 \geq 0 \end{aligned} \tag{12.173}$$

These non-negative values a_1, a_2 and a_3 can be evaluated by solving the following system of linear equations by crammer's rule of determinants

$$\begin{aligned} \mu_1^{(left)}a_1 + \mu_1^{(left)}a_2 + \mu_1^{(right)}a_3 &= \mu_1^* \\ \mu_2^{(S)}a_1 + \mu_2^{(P)}a_2 + \mu_2^{(Q)}a_3 &= \mu_2^* \\ a_1 + a_2 + a_3 &= 1 \end{aligned} \tag{12.174}$$

Therefore, with subject to $\mu_2^{(S)} > \mu_2^{(P)}$ and $\mu_2^{(right)} > \mu_2^{(left)}$, we get the solution of (12.174) as follows:

$$\begin{aligned} a_1 &= \frac{\begin{vmatrix} \mu_1^{(left)} & \mu_2^{(P)} & 1 \\ \mu_1^{(right)} & \mu_2^{(Q)} & 1 \\ \mu_1^* & \mu_2^* & 1 \end{vmatrix}}{\left(\mu_2^{(S)} - \mu_2^{(P)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} = \frac{CCW(P,Q,T)}{\left(\mu_2^{(S)} - \mu_2^{(P)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} \\ &\geq 0, \text{ since } T \text{ is on the left of } \overrightarrow{PQ} \text{ or on } \overline{PQ} \text{ itself} \end{aligned} \tag{12.175}$$

$$\begin{aligned} a_2 &= \frac{\begin{vmatrix} \mu_1^{(right)} & \mu_2^{(Q)} & 1 \\ \mu_1^{(left)} & \mu_2^{(S)} & 1 \\ \mu_1^* & \mu_2^* & 1 \end{vmatrix}}{\left(\mu_2^{(S)} - \mu_2^{(P)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} = \frac{CCW(Q,S,T)}{\left(\mu_2^{(S)} - \mu_2^{(P)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} \\ &> 0, \text{ since } T \text{ is on the left of } \overrightarrow{QS} \end{aligned} \tag{12.176}$$

$$\begin{aligned} a_3 &= \frac{\begin{vmatrix} \mu_1^{(left)} & \mu_2^{(S)} & 1 \\ \mu_1^{(left)} & \mu_2^{(P)} & 1 \\ \mu_1^* & \mu_2^* & 1 \end{vmatrix}}{\left(\mu_2^{(S)} - \mu_2^{(P)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} = \frac{CCW(S,P,T)}{\left(\mu_2^{(S)} - \mu_2^{(P)}\right)\left(\mu_1^{(right)} - \mu_1^{(left)}\right)} \\ &> 0, \text{ since } T \text{ is on the left of } \overrightarrow{SP} \end{aligned} \tag{12.177}$$

If T is found to lie exactly on $\overline{PQ}$, it is clear, that $\overline{PQ}$ must be parallel to the horizontal axis of the input plane, where $\mu_2^{(P)} = \mu_2^{(Q)}$.

As a result of the solution of (12.174), by using the selected records $_{(\gamma^{(S)},\beta^{(S)},\mu_1^{(left)},\mu_2^{(S)})}$, $_{(\gamma^{(P)},\beta^{(P)},\mu_1^{(left)},\mu_2^{(P)})}$ and $_{(\gamma^{(Q)},\beta^{(Q)},\mu_1^{(right)},\mu_2^{(Q)})}$, the access vector is given as (β_a, γ_a), such that

$$\begin{aligned} \beta_a &= a_1\beta^{(S)} + a_2\beta^{(P)} + a_3\beta^{(Q)} \\ \gamma_a &= a_1\gamma^{(S)} + a_2\gamma^{(P)} + a_3\gamma^{(Q)} \end{aligned} \tag{12.178}$$

Therefore, the consideration of the entire case $\left(\mu_2^{(Q)} > \mu_2^{(P)}\right)$ comes to the end, where **four** records are needed to be accessed for choosing our access vector

Hence, we have covered both the cases, namely $\mu_2^{(Q)} = \mu_2^{(P)}$ and $\mu_2^{(Q)} > \mu_2^{(P)}$, where the procedures for yielding the access vector (β_a, γ_a) are different.

The access vector is therefore just the starting point of the ultimate solution (β^*, γ^*) immediately after the database access. However, the question of it's compatibility for the Newton Raphson procedure or even for the iterative procedure still remains open.

It requires subsequent processing for the sake of the improvement of it's compatibility .

12.7.7 Algorithm for subsequent processing of access vectors

The execution of the subsequent processing of the access vector (β_a, γ_a) is actually meant for the cases, when $\mu_1^{(Q)} > \mu_1^{(P)}$ and at the same time $\left((\mu_1^*, \mu_2^*) \in \Delta SRQ \setminus \overline{SQ}\right)$ or $\left((\mu_1^*, \mu_2^*) \in \Delta SPQ \setminus \overline{SQ}\right)$.

The idea will be to make subsequent individual considerations of the three smaller triangles having $T : (\mu_1^*, \mu_2^*)$ as their common vertex, which came into being as a result of the subdivision of the triangle ΔSRQ or ΔSPQ according as T belongs to $\Delta SRQ \setminus \overline{SQ}$ or $\Delta SPQ \setminus \overline{SQ}$.

For the purpose of the algorithmic construction, the relevant points of both the triangles ΔSRQ and ΔSPQ must necessarily be renamed. That is,

- in ΔSRQ, the points S, R , Q and T are renamed as A, B , C and D respectively. The diagrammatic representation of ΔSRQ is given as

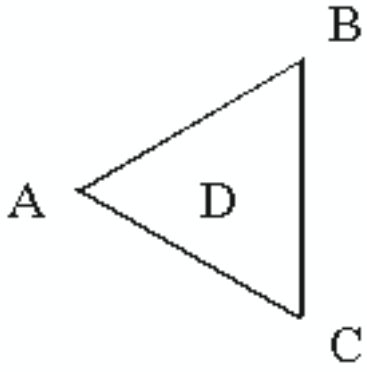


Figure 12.5: Renamed ΔSRQ

- in ΔSPQ, the points S, P , Q and T are renamed as A, B , C and D respectively. The diagrammatic representation of ΔSPQ is given as

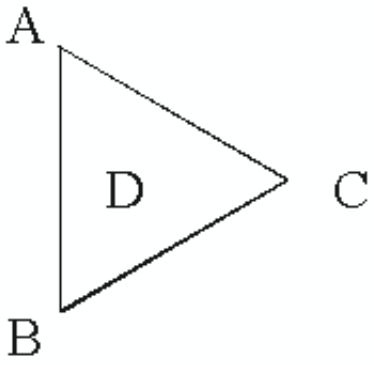


Figure 12.6: Renamed ΔSPQ

Till now, in order to yield the access vector (β_a, γ_a), the $det_{\overrightarrow{SQ}} = CCW(S, Q, T)$ has already been computed by means of (12.161) in order to reach one of the

following conclusions: $T \in \Delta SRQ \setminus \overline{SQ}$ or $T \in \Delta SPQ \setminus \overline{SQ}$ according as $det_{\overrightarrow{SQ}} \gtrless 0$.

Immediately on finding the access vector (β_a, γ_a), as already described in the previous subsection, the process of renaming the points P, Q, R and S is carried out in the following way

- if $det_{\overrightarrow{SQ}} > 0$, then S, R and Q are renamed as A, B and C respectively as shown in the figure 12.5
- if $det_{\overrightarrow{SQ}} < 0$, then S, P and Q are renamed as A, B and C respectively as shown in the figure 12.6

After this, the values of $\mu_1^{(\beta_a,\gamma_a)}$ and $\mu_2^{(\beta_a,\gamma_a)}$, which are subjected to the definitions given in (12.85) and (12.86) respectively, are computed by means of (12.107) or (12.110).

Then, it is examined, whether (β_a, γ_a) fulfills both the convergence conditions (12.101) and (12.102) simultaneously or not. Computations of the moments of higher orders (up to the moments of sixth order) in this regard are carried out by means of (12.107) or (12.110) successively. Therefore,

- If these conditions are fulfilled, then the execution of this procedure is not necessary anymore and the current access vector is forwarded for the next procedure
- Otherwise, then the execution of this procedure needs to be started by choosing the point D to be $D : \left(\mu_1^{(\beta_a,\gamma_a)}, \mu_2^{(\beta_a,\gamma_a)}\right)$ instead of $D : (\mu_1^*, \mu_2^*)$. This is done by replacing it's previously defined coordinates by the newly computed coordinates.

  The basic idea of choosing this new D is, that the image of the new point $D : \left(\mu_1^{(\beta_a,\gamma_a)}\mu_2^{(\beta_a,\gamma_a)}\right)$ in the input space, which is (β_a, γ_a) in the solution space, is known at the very moment.

  Of course, our ultimate aim (our aim in the long run) is to determine the image of (μ_1^*, μ_2^*) in the solution space.

Before, we proceed further, we initialize our start vector (β_s, γ_s) with the help of the existing access vector (β_a, γ_a) as

- $\beta_s := \beta_a$ and $\gamma_s := \gamma_a$

Thereafter, the final part of this procedure is programmed within a *while* loop, which is broken, if any one of the following situations is reached:

- The start vector (β_s, γ_s) fulfills both the convergence conditions (12.101) and (12.102) simultaneously at the end of any cycle
- The number of cycles of the *while* loop exceeds 10
- A given conditional breakage of the loop, if T is found to lie on one of the segments $\overline{AD}$, $\overline{BD}$ or $\overline{CD}$ at the end of any cycle

Therefore, keeping in mind, that our start vector (β_s, γ_s), which turns up as a result of the processing of the access vector feeded to the procedure as the input, the *while* loop containing the rest of the desired procedure is described as
while(none of the above three stated situations are reached)
{

1. Compute $CCW(A, D, T)$, $CCW(B, D, T)$ and $CCW(C, D, T)$ by the rule (12.156)
2. If $((CCW(A, D, T) = 0) \vee (CCW(B, D, T) = 0) \vee (CCW(C, D, T) = 0))$ then as the case may be, i.e. $V \in \{A, B, C\}$,

   {

   - $T : (\mu_1^*, \mu_2^*)$ is expressed as a linear combination of A and V as described in (12.168)
   - the output start vector (β_s, γ_s) is derived exactly in the identical manner as described in (12.172)
   - the *while* loop must be **broken** at this point with the result of this procedure as (β_s, γ_s).

     This necessarily meant, that T lies on one of the segments $\overline{AD}$, $\overline{BD}$ or $\overline{CD}$

   }

3. If otherwise, (i.e. if T is not found to lie on one of the segments $\overline{AD}$, $\overline{BD}$ or $\overline{CD}$)
   then it is clear, that T must lie within only one of the triangles ΔABD, ΔBCD or ΔCAD (i.e. no two of these triangles can contain T jointly), then

   {

   - By the rule (12.156), compute the following
     - $CCW(A, B, T)$, $CCW(B, D, T)$, $CCW(D, A, T)$
     - $CCW(B, C, T)$, $CCW(C, D, T)$, $CCW(D, B, T)$
     - $CCW(C, A, T)$, $CCW(A, D, T)$, $CCW(D, C, T)$

   - if $det_{\overrightarrow{SQ}} > 0$, then
     - if
       $(CCW(A, B, T)<0 \wedge CCW(B, D, T)<0 \wedge CCW(D, A, T)<0)$
       then, since T must lie within the triangle ΔABD,

       {
       * $T : (\mu_1^*, \mu_2^*)$ is expressed as a convex combination of A, B and D as described in (12.162)
       * The point D is renamed as C and thereby the ΔABD is renamed as ΔABC
       * The vector (β_s, γ_s) is derived exactly in the identical manner as described in (12.167) and is named as D
       * The values of the moments up to the sixth order, i.e. $\mu_n^{(\beta_s,\gamma_s)}$ $(n \in \{1, 2, \ldots, 6\})$, are computed successively by means of (12.107) or (12.110)
       * Then, it is examined, whether (β_s, γ_s) fulfills both the convergence conditions (12.101) and (12.102) simultaneously

       }

- if
  $(CCW(B,C,T)<0 \wedge CCW(C,D,T)<0 \wedge CCW(D,B,T)<0)$
  then, since T must lie within the triangle ΔBCD,

  {
  - $T:(\mu_1^*,\mu_2^*)$ is expressed as a convex combination of B, C and D as described in (12.162)
  - The points B, C and D are renamed as A, B and C respectively and thereby the ΔBCD is renamed as ΔABC
  - The vector (β_s,γ_s) is derived exactly in the identical manner as described in (12.167) and is named as D
  - The values of the moments up to the sixth order, i.e. $\mu_n^{(\beta_s,\gamma_s)}$ $(n \in \{1,2,\ldots,6\})$, are computed successively by means of (12.107) or (12.110)
  - Then, it is examined, whether (β_s,γ_s) fulfills both the convergence conditions (12.101) and (12.102) simultaneously

  }
- if
  $(CCW(C,A,T)<0 \wedge CCW(A,D,T)<0 \wedge CCW(D,C,T)<0)$
  then, since T must lie within the triangle ΔCAD,

  {
  - $T:(\mu_1^*,\mu_2^*)$ is expressed as a convex combination of C, A and D as described in (12.162)
  - The points C, A and D are renamed as A, B and C respectively and thereby the ΔCAD is renamed as ΔABC
  - The vector (β_s,γ_s) is derived exactly in the identical manner as described in (12.167) and is named as D
  - The values of the moments up to the sixth order, i.e. $\mu_n^{(\beta_s,\gamma_s)}$ $(n \in \{1,2,\ldots,6\})$, are computed successively by means of (12.107) or (12.110)
  - Then, it is examined, whether (β_s,γ_s) fulfills both the convergence conditions (12.101) and (12.102) simultaneously

  }

- if $det_{\overrightarrow{SQ}} < 0$, then
  - if $(CCW(A,B,T){>}0 \wedge CCW(B,D,T){>}0 \wedge CCW(D,A,T){>}0)$ then, since T must lie within the triangle ΔABD,

    {
    * $T : (\mu_1^*, \mu_2^*)$ is expressed as a convex combination of A, B and D as described in (12.173)
    * The point D is renamed as C and thereby the ΔABD is renamed as ΔABC
    * The vector (β_s, γ_s) is derived exactly in the identical manner as described in (12.178) and is named as D
    * The values of the moments up to the sixth order, i.e. $\mu_n^{(\beta_s,\gamma_s)}$ $(n \in \{1, 2, \ldots, 6\})$, are computed successively by means of (12.107) or (12.110)
    * Then, it is examined, whether (β_s, γ_s) fulfills both the convergence conditions (12.101) and (12.102) simultaneously

    }
  - if $(CCW(B,C,T){>}0 \wedge CCW(C,D,T){>}0 \wedge CCW(D,B,T){>}0)$ then, since T must lie within the triangle ΔBCD,

    {
    * $T : (\mu_1^*, \mu_2^*)$ is expressed as a convex combination of B, C and D as described in (12.173)
    * The points B, C and D are renamed as A, B and C respectively and thereby the ΔBCD is renamed as ΔABC
    * The vector (β_s, γ_s) is derived exactly in the identical manner as described in (12.178) and is named as D
    * The values of the moments up to the sixth order, i.e. $\mu_n^{(\beta_s,\gamma_s)}$ $(n \in \{1, 2, \ldots, 6\})$, are computed successively by means of (12.107) or (12.110)
    * Then, it is examined, whether (β_s, γ_s) fulfills both the convergence conditions (12.101) and (12.102) simultaneously

    }

- if
  $(CCW(C,A,T)>0 \wedge CCW(A,D,T)>0 \wedge CCW(D,C,T)>0)$
  then, since T must lie within the triangle ΔCAD,

  {
  - $T:(\mu_1^*,\mu_2^*)$ is expressed as a convex combination of C, A and D as described in (12.173)
  - The points C, A and D are renamed as A, B and C respectively and thereby the ΔCAD is renamed as ΔABC
  - The vector (β_s,γ_s) is derived exactly in the identical manner as described in (12.178) and is named as D
  - The values of the moments up to the sixth order, i.e. $\mu_n^{(\beta_s,\gamma_s)}$ $(n \in \{1,2,\ldots,6\})$, are computed successively by means of (12.107) or (12.110)
  - Then, it is examined, whether (β_s,γ_s) fulfills both the convergence conditions (12.101) and (12.102) simultaneously

  }

}

}

This is the end of the execution of the *while* loop and thereby the end of the desired procedure for the first stage processing of the access vector. The output delivered by the procedure is the vector (β_s,γ_s).

It has to be noted, that the triangle ΔABC containing the point $D:(\beta_s,\gamma_s)$ becomes gradually smaller after each cycle of the *while* loop. At the end of each cycle, (β_s,γ_s) comes a step undoubtedly closer to the solution (β^*,γ^*). However, the maximum number of cycles is limited to ten, within which (β_s,γ_s) can fulfill the necessary convergence conditions. Fulfillment of these convergence conditions for the sake of the compatibility of (β_s,γ_s) for the Newton Raphson procedure is the core idea of this procedure, since the compatibility of (β_a,γ_a) for the same was still a question. In the bulk of the cases, this fulfillment takes place. Only in cases, when the user-given μ_2^* has extreme values (i.e. values closer to either μ_1^* or μ_1^{*2}), this fulfillment may not take place within ten cycles. But this is not a serious problem. The modified iterative procedures introduced in the immediately next subsection handles these cases.

Therefore, we conclude, that (β_s, γ_s) is not the start vector, which can be forwarded to Newton Raphson procedure right at this stage. This vector must necessarily be processed further to ensure the perfect functioning of the Newton Raphson procedure. This processing handled by the two modified iterative procedures, which shall be discussed in the next subsection, shall be the second stage processing or rather the final stage processing before the Newton Raphson procedure.

In this procedure, the programmer has wilfully limited the number of cycles of the *while* loop to ten, because this is a time consuming procedure in certain cases and total allotted time meant for running the entire program was always needed to be kept in mind.

As a matter of fact, the compatibility of the access vector for the Newton Raphson procedure arising out of the cases, when $\mu_1^{(Q)} = \mu_1^{(P)}$ never seemed to be a big concern.

However, it should not be assumed, that the cases involving $\mu_1^{(Q)} > \mu_1^{(P)}$ and $(\mu_1^*, \mu_2^*) \in \overline{SQ}$ are not within consideration. The modified iterative procedures discussed in the immediately next subsection handles such cases anyway.

12.7.8 Modified iterative procedures

Referring to the subsection *Characteristics of the outputs with respect to the inputs*, it is already known, the proper choice of a suitable start vector (β_s, γ_s), in cases when $\mu_2 > \mu_{2upp}$ or $\mu_2 < \mu_{2low}$ is difficult in general.

The introduction of the modified iterative procedures is essentially meant for cases, when the user-given variances are too large or too small.

In other words, since the extreme largeness and extreme smallness of the user-given variances are directly linked with the cases, when $\mu_2 > \mu_{2upp}$ or $\mu_2 < \mu_{2low}$ respectively, modified iterative procedures are the suitable procedural measures to ensure the proper choice of a start vector (β_s, γ_s)

The modified iterative procedures are no less useful, even for cases, when $\mu_{2low} \leq \mu_2 \leq \mu_{2upp}$. They can only improve a given start vector in terms of it's closeness to the solution (β^*, γ^*) and never the opposite.

The two modified iterative procedures are distinguished by the two cases, namely $\mu_2 > \mu_{2upp}$ and $\mu_2 < \mu_{2low}$. Both these procedures are somewhat similar to the designed iterative procedure (12.91), but changed in certain ways conveniently.

Therefore, the algorithms for both the modified iterative procedures are given by

Both the following introduced procedures begin with the following pre-computations,

- $e\_distance$ (as described in (12.88)) is computed with respect to the values of β_s and γ_s by computing the values of $\mu_1^{(\beta_s,\gamma_s)}$ and $\mu_2^{(\beta_s,\gamma_s)}$ by means of (12.107) or (12.110)

- we initialize $\beta = \beta_s$, $\gamma = \gamma_s$, $\mu_1 = \mu_1^{(\beta_s,\gamma_s)}$ and $\mu_2 = \mu_2^{(\beta_s,\gamma_s)}$ together with $becomes\_smaller = true$

- a suitably chosen fixed integer $n \in \mathbb{Z}$

- a small positive number ϵ is suitably chosen

Therefore, with subject to the above data,

1. Specially for ($\mu_2 > \mu_{2upp}$) cases:

   while $((\gamma > 10^n) \wedge (e\_distance > \epsilon) \wedge becomes\_smaller)$
   {
   $\gamma_{reserve} = \gamma$; $\beta_{reserve} = \beta$; $\mu_{1reserve} = \mu_1$; $\mu_{2reserve} = \mu_2$;
   $\gamma- = 10^n$; (means, γ is decremented by 10^n)
   $\beta = SolveForBeta(\gamma, \beta)$; (as described in (12.89))

   Compute the values of $\mu_1^{(\beta,\gamma)}$ and $\mu_2^{(\beta,\gamma)}$ by means of (12.107) or (12.110) with subject to the newly updated values of β and γ. Then update the values of μ_1 and μ_2 as $\mu_1 = \mu_1^{(\beta,\gamma)}$ and $\mu_2 = \mu_2^{(\beta,\gamma)}$;

   $e\_distance\_current$ = Computed $e\_distance$ with respect to β and γ;

   if ($e\_distance \geq e\_distance\_current$) then $e\_distance = e\_distance\_current$

   else
   {
   $becomes\_smaller = false$;
   (means, deterioration of the start vector & while loop must be broken immediately on resetting with the previous values in the following step)

   $\gamma = \gamma_{reserve}$; $\beta = \beta_{reserve}$; $\mu_1 = \mu_{1reserve}$; $\mu_2 = \mu_{2reserve}$;
   }

   }

   $becomes\_smaller = true$;
   Set $\gamma_s = \gamma$; $\beta_s = \beta$;
   (β_s, γ_s) is the improved start vector as a result of the procedure

(12.179)

2. Specially for $(\mu_2 < \mu_{2low})$ cases:

   while $((\gamma < 10^{-n}) \wedge (e\_distance > \epsilon) \wedge becomes\_smaller)$
   {
   $\gamma_{reserve} = \gamma; \quad \beta_{reserve} = \beta; \quad \mu_{1reserve} = \mu_1; \quad \mu_{2reserve} = \mu_2;$
   $\gamma += 10^n$; (means, γ is incremented by 10^n)
   $\beta = SolveForBeta(\gamma, \beta)$; (as described in (12.89))

   Compute the values of $\mu_1^{(\beta,\gamma)}$ and $\mu_2^{(\beta,\gamma)}$ by means of (12.107) or (12.110) with subject to the newly updated values of β and γ. Then update the values of μ_1 and μ_2 as $\mu_1 = \mu_1^{(\beta,\gamma)}$ and $\mu_2 = \mu_2^{(\beta,\gamma)}$;

   $e\_distance\_current$ = Computed $e\_distance$ with respect to β and γ;

   if $(e\_distance \geq e\_distance\_current)$ then $e\_distance = e\_distance\_current$

   else
   {
   $becomes\_smaller = false$;
   (means, deterioration of the start vector & while loop must be broken immediately on resetting with the previous values in the following step)
   $\gamma = \gamma_{reserve}; \quad \beta = \beta_{reserve}; \quad \mu_1 = \mu_{1reserve}; \quad \mu_2 = \mu_{2reserve};$
   }

   }

   $becomes\_smaller = true$;
   Set $\gamma_s = \gamma$; $\beta_s = \beta$;
   (β_s, γ_s) is the improved start vector as a result of the procedure

   (12.180)

These modified iterative procedures help to reduce the running time of the program considerably.

12.7.9 The subroutine for non special uni-extremal cases

The solution of the system of equations (12.130), where X is a continuous random variable with the range of variability $\{x|0 \leq x \leq 1\}$, is identical with the solution of the system of equations (12.94) for the continuous case. This justifies the usage of the designed Newton Raphson numerical procedure (12.115) for the solution of the system (12.130) as well.

Therefore, the algorithm for non special cases, which aims to solve the system (12.130) step by step, is described by the following sequential steps:

1. At the very first step of the execution of the program, the $Z = 1 - X$ transformation is carried out, if the user-given input (μ_1^*, μ_2^*) lies within the specified areas of the input space described in (12.146).

   In such cases, the program shall work with the moments of Z as it's input instead of the moments of X and therefore the yielded (β_0, γ_0) before being processed by the Newton Raphson procedure will become the final first approximated solution of the equation system (12.151) instead of the equation system (12.130).

2. if $\mu_1^* < 0.01$ or $\mu_1^* > 0.99$, then the determination of the final start vector (β_0, γ_0) meant for forwarding to the Newton Raphson procedure is not possible by database access, but only by the iterative procedure (12.91). However, the running time of the program is well beyond the programmer's control and therefore beyond any estimation.

   In this case, we go by the following steps:

   - If $\mu_1^* < 0.01$, then solve the following equation (in β) by the solution procedure (12.36) after resetting the equation into $f(\beta) = 0$ form:
   $$\mu^* = \begin{cases} 0.5 & : \quad \beta = 0 \\ 1 + \frac{1}{e^\beta - 1} - \frac{1}{\beta} & : \quad \beta \neq 0 \end{cases}$$
   Let β_s be the yielded solution
   - If $\mu_1^* > 0.99$, then since the usage of the $Z = 1 - X$ transformation is necessary, solve the following equation (in β) by the solution procedure (12.36) after resetting the equation into $f(\beta) = 0$ form:
   $$1 - \mu^* = \begin{cases} 0.5 & : \quad \beta = 0 \\ 1 + \frac{1}{e^\beta - 1} - \frac{1}{\beta} & : \quad \beta \neq 0 \end{cases}$$

Let β_s be the yielded solution

- We initialize $\beta = \beta_s$ and $\gamma = 0$
- Compute the final start vector (β_0, γ_0) by the iterative procedure (12.91) with subject to $\epsilon = 10^{-5}$. Immediately after this, go to the step (8.) for the Newton Raphson procedure

3. if $0.01 \leq \mu_1^* \leq 0.99$, then the database will be accessed within the specified area (12.155) of the input space. Here, the values of ℓ ranging from 2.08699 to 3.5 have been chosen carefully by the programmer to ensure the yield of the best possible results in terms of their accuracies.

4. Compute the access vector (β_a, γ_a) on accessing 1, 2 or 4 records from the database (as the case may be), as discussed

5. Compute the improved start vector (β_s, γ_s), if the access vector (β_a, γ_a) can be processed subsequently (i.e. in cases for $det_{\overrightarrow{SQ}} > 0$ or $det_{\overrightarrow{SQ}} < 0$)

6. Using this start vector (β_s, γ_s) or the access vector (β_a, γ_a) (in case a further processing was not possible), execute the modified iterative procedure (12.179) with respect to $n = 3, 2, 1, 0, -1$ successively to yield an even better (*improved*) start vector (β_s, γ_s).

   The program-control recognizes the need of this procedure and executes it intensively when the value of μ_2^* is rather high.

   Immediately after this, the control uses this (β_s, γ_s) to execute the iterative procedure (12.91) with respect to $\epsilon = 10^{-3}$ for a maximum number of 6 iterations, provided $\gamma_s > 0$ still holds

7. In this step, the cases will be considered, when either both the user-given mean (μ_1^*) and the user-given variance $(\sigma^2 = \mu_2^* - \mu_1^{*2})$ are relatively small or σ^2 is small but not enough to ensure the resulting probability density function of Y denoted by $f_{Y|\{d_Y\}}(y)$ to be accepted as the expression given in (12.136). These cases arise, when either of the following happens:

   - $0.01 < \mu_1^* \leq 0.025$ and $\sigma^2 < 0.0012$
   - $0.025 < \mu_1^* < 0.05$ and $\sigma^2 < 0.001$
   - $\gamma_s < -1000$

- $0.01 < \mu_1^* < 0.05$ and $\gamma_s < 0$

In either of the above cases, execute the modified iterative procedure (12.180) with respect to $n = 5, 3, 0.69879, 0$ successively to yield an even better (*improved*) start vector (β_s, γ_s).

The program-control recognizes the need of this procedure and executes it intensively when the value of μ_2^* is rather low.

Obviously, $n = 0.69879$ was made to be a slight exception to the modified iterative procedural rule (12.180), where $n \notin \mathbb{Z}$. This solely means, that γ has been incremented by 5 in each cycle of the while loop.

Immediately after this, the control uses this (β_s, γ_s) to execute the iterative procedure (12.91) with respect to $\epsilon = 10^{-5}$. This procedure is continued till the *e_distance* is reduced at the very least.

8. The Newton Raphson procedure described by (12.115) is executed with subject to

   - $\epsilon = 10^{-16}$
   - the number of *additional criteria* are restricted to one only. This criterion is, that the maximum number of Newton Raphson procedural cycles is limited to 10000
   - the first approximated solution (β_0, γ_0), which is
     - of the equation system (12.130), if the program has to work with the moments of X
     - of the equation system (12.151), if the program has to work with the moments of Z

   As a result of the execution of the Newton Raphson procedure, the final solution of the equation system (12.130) is yielded as (β^*, γ^*) or the same of the equation system (12.151) is yielded as $(\overline{\beta^*}, \overline{\gamma^*})$ according as the program had worked with the moments of X or of Z

9. The reverse transformation, namely $X = 1 - Z$ will be necessary, if the program had worked with the moments of Z.

   In that case, with the help of $(\overline{\beta^*}, \overline{\gamma^*})$, final desired solution (β^*, γ^*) shall be computed as described in (12.154)

Thus, the computed (β^*, γ^*) in the final step is the desired solution of (12.130).

Whence, by using (6.75) too, $f_{Y|\{d_Y\}}(y)$ is finally expressed as

$$f_{Y|\{d_Y\}}(y) = \frac{1}{b-a} \frac{e^{\beta^*\left(\frac{y-a}{b-a}\right)+\gamma^*\left(\frac{y-a}{b-a}\right)^2}}{\int\limits_0^1 e^{\beta^* t+\gamma^* t^2} dt} = \frac{K}{\sigma_Y\sqrt{2\pi}} e^{\widehat{\lambda}\left(\frac{y-M}{\sigma_Y}\right)^2}, \;\; a \leq y \leq b \tag{12.181}$$

12.8 Limitations

12.8.1 Probable limitations of my software programs in monotone cases

In this subsection, we state briefly, for which input parameters both the software programs (referred to the discrete and continuous monotone cases) **may not** be successfully executable.

The program may not deliver an output, if $\mu_Y^{(1)}$ happens to be too large or too small, i.e. if $\mu_Y^{(1)}$ is chosen to be too close to a from right or too close to b from left.

From the **stochastic point of view**, such inputs may legitimately taken for **simple degenerated probability distributions**, as already discussed.

12.8.2 Probable limitations of my software program in discrete uni-extremal cases

In this subsection, we state briefly, for which input parameters the software program (referred to the discrete uni-extremal cases) **may not** be successfully executable.

The program may not deliver an output, if $\mu_Y^{(1)}$ happens to be too large or too small, i.e. if $\mu_Y^{(1)}$ is chosen to be too close to a from right or too close to b from left. In other words, the user **may** think of **not** choosing $\mu_Y^{(1)}$ in one of the following ways:

- $a < \mu_Y^{(1)} < a + 0.01(b-a)$
- $b > \mu_Y^{(1)} > b - 0.01(b-a)$

Of course, such cases of failure are **extremely rare**. Such inputs may not be legitimately justified either, especially from the **stochastic point of view**.

12.8.3 Limitations of my software program in continuous uni-extremal cases

In this subsection, we state briefly, for which input parameters the software program (referred to the continuous uni-extremal cases) cannot be properly

executed. Of course, such cases, when the input parameters are not acceptable by the software program, are extremely few.

We shall cite such cases one by one as follows:

1. The program does not give the desired results, if $\mu_Y^{(2)}$ (or equivalently the variance $\sigma_Y^2 = \mu_Y^{(2)} + \left(\mu_Y^{(1)}\right)^2$ can be given as input) is chosen too close to $\mu_Y^{(1)}$ from left, after $\mu_Y^{(1)}$ has been input by the user.

   As for examples,

   (a) for $a = 4.5$, $b = 10.7$, $\mu_Y^{(1)} = 6.6$, $\mu_Y^{(2)}$ must be theoretically restricted to $43.56 < \mu_Y^{(2)} < 52.17$. If $\mu_Y^{(2)} = 52.1$ is given as input, the program does not deliver the output result, because for this particular high value of $\mu_Y^{(2)}$ the database cannot deliver the correct records for the computation of the start vector. However, this problem gets fruitfully resolved, if $\mu_Y^{(2)}$ is chosen to be $\mu_Y^{(2)} = 51.5$ instead.

   (b) for $a = 20.4$, $b = 55.48$, $\mu_Y^{(1)} = 42.62$, σ_Y^2 must be theoretically restricted to $0 < \sigma_Y^2 < 285.7492$. If $\sigma_Y^2 = 285.2$ is given as input, the program does not deliver the output result for the same reason as stated immediately above. However, this problem gets fruitfully resolved, if σ_Y^2 is chosen to be $\sigma_Y^2 = 284.5$ instead.

2. The execution of the program can take unusually or even indefinitely long time, if $\mu_Y^{(1)}$ is chosen to be too close to either a or b. Precisely, this happens, if either $a < \mu_1^{(1)} < a + 0.01(b-a)$ or $b - 0.01(b-a) < \mu_1^{(1)} < b$ is the case.

   However, this problem gets resolved, if $\mu_Y^{(2)}$ (or equivalently σ_Y^2) is small enough for the delivery of appropriate approximative results with the help of the standard normal distribution (as already discussed). As for eg. for $a = 20.4$, $b = 55.48$, $\mu_Y^{(1)} = 20.6$, the programm delivers the result for $\sigma_Y^2 = 0.001$ quickly. A higher value of σ_Y^2 can be a big problem, as far as the time of the program execution is concerned

12.9 Operating instructions

The **four** software programs developed in the **object oriented programming language, Java** can be run in any personal computer, where **Java** is installed. These four software programs have already been compiled. Both the source files and the compiled object (class) files and burnt into the supplied CD ROM.

The main folder burnt in the CD ROM named **SurathTextMIP** contains another folder named **ProjektInformal**. This folder ProjektInformal contains five folders, whose description are given as follows:

1. The folder **auxiliarydata**: This folder contains all the tables in form of text files that are accessible by particular java class file. These tables contain the necessary starting values for running one of the four the developed java programs.

2. The folder **bin** contains all the **compiled .class files** that are related to the two java programs involving the **continuous random variable** Y (i.e. when the random variable Y is continuous).

3. The folder **binDiscrete** contains all the **compiled .class files** that are related to the two java programs involving the **discrete random variable** Y (i.e. when the random variable Y is discrete).

4. The folder **Sources** contains all the java classes, i.e. the complete java source code in form of .java files involving the **continuous random variable** Y.

5. The folder **SourcesDiscrete** contains all the java classes, i.e. the complete java source code in form of .java files involving the **discrete random variable** Y.

If the compiled class files burnt in the CD ROM match the java-compiler installed in the PC, then the software programs can be run on the CD ROM itself in the PC automatically. **If this is not the case**, then the entire folder **SurathTextMIP** needs to be copied into the PC and the user needs to recompile the java source programs thereafter before running them.

The compilation and the running of the java source programs are described as follows:

- For the computation of a monotonic discrete probability distribution of Y, the java class **MonotonicDiscrete.java** needs to be
  - (re)compiled by double clicking the batch file **compileMonotonicDiscrete** situated in SurathTextMIP\ProjektInformal**sourcesDiscrete**
  - run by double clicking the batch file **runMonotonicDiscrete** situated in SurathTextMIP\ProjektInformal**binDiscrete**
- For the computation of a uni-extremal discrete probability distribution of Y, the java class **UniExtremalDiscrete.java** needs to be
  - (re)compiled by double clicking the batch file **compileUniExtremalDiscrete** situated in SurathTextMIP\ProjektInformal**sourcesDiscrete**
  - run by double clicking the batch file **runUniExtremalDiscrete** situated in SurathTextMIP\ProjektInformal**binDiscrete**
- For the computation of a monotonic continuous probability distribution of Y, the java class **MonotonicContinuous.java** needs to be
  - (re)compiled by double clicking the batch file **compileMonotonicContinuous** situated in SurathTextMIP\ProjektInformal**sources**
  - run by double clicking the batch file **runMonotonicContinuous** situated in SurathTextMIP\ProjektInformal**bin**
- For the computation of a uni-extremal continuous probability distribution of Y, the java class **UniExtremalContinuousTEXT.java** needs to be
  - (re)compiled by double clicking the batch file **compileUniExtremalContinuous** situated in SurathTextMIP\ProjektInformal**sources**

– run by double clicking the batch file
runUniExtremalContinuous situated in
SurathTextMIP\ProjektInformal**bin**

Appendix A

The maximization of the stochastic entropy

A.1 The expression of the maximum entropy probability distribution

Referring to the subsection 3.3.1, we shall use identically the same notations in this section. In fact, for the sake of a brief restatement, let $\mathcal{F}(m)$ be the set of all the probability distributions $\mathbf{P}_{Y|\{d_Y\}}$ on $(\Omega, \mathcal{A})$, such that each of the elements of $\mathcal{F}(m)$, namely $\mathbf{P}_{Y|\{d_Y\}}$ possesses a density $\mathbf{f}_{Y|\{d_Y\}}$ with respect to ν and has the k th moment equal to a fixed number $\mu_Y^{(k)}$, for $k = 0, 1, 2, \ldots, m$. Here, $d_Y = (\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(m)})$.

In this section, we shall show that the exponential polynomial probability distribution (denoted by $\mathbf{P}_{Y|\{d_Y\}}^{MEP}$) of the random variable Y determined uniquely by its first m moments $\mu_Y^{(k)}$, $k = 0, 1, 2, \ldots, m$, has the **maximum entropy** among all other probability distributions having identically the **same support** as well as identically the **same first** m **moments**.

This proof, which is the **theorem 12.1.1** given in the page 410 of the book [11], is basically given by means of **Kullback-Leibler divergence** (or **relative entropy**). This proof is given as follows:

Proof of the aforesaid ***theorem 12.1.1.*** *i.e. the proof of the* ***theorem 3.3.1****.* Clearly, $\mathbf{P}_{Y|\{d_Y\}}^{MEP} \in \mathcal{F}(m)$. We shall make use of the very fact that the rela-

tive entropy $\int\limits_{\Omega} \mathbf{f}_{Y|\{d_Y\}} \log\left(\frac{\mathbf{f}_{Y|\{d_Y\}}}{f_{Y|\{d_Y\}}}\right) d\nu$ (i.e. the entropy of $\mathbf{P}_{Y|\{d_Y\}}$ relative to $\mathbf{P}^{MEP}_{Y|\{d_Y\}}$) is always nonnegative. This nonnegativity is referred to the **Gibbs' inequality**.

For any arbitrarily chosen $\mathbf{P}_{Y|\{d_Y\}} \in \mathcal{F}(m)$ with a ν- density $\mathbf{f}_{Y|\{d_Y\}}$, we have

$$
\begin{aligned}
& H\left(\mathbf{P}_{Y|\{d_Y\}}\right) - H\left(\mathbf{P}^{MEP}_{Y|\{d_Y\}}\right) \\
= & \int\limits_{\Omega} f_{Y|\{d_Y\}} \log\left(f_{Y|\{d_Y\}}\right) d\nu - \int\limits_{\Omega} \mathbf{f}_{Y|\{d_Y\}} \log\left(\mathbf{f}_{Y|\{d_Y\}}\right) d\nu \\
= & \int\limits_{\Omega} \left(f_{Y|\{d_Y\}} \log\left(f_{Y|\{d_Y\}}\right) - \mathbf{f}_{Y|\{d_Y\}} \log\left(f_{Y|\{d_Y\}}\right) \right) d\nu \\
& \underbrace{- \int\limits_{\Omega} \mathbf{f}_{Y|\{d_Y\}} \log\left(\frac{\mathbf{f}_{Y|\{d_Y\}}}{f_{Y|\{d_Y\}}}\right) d\nu}_{\geq 0} \qquad (\text{ Gibbs' Inequality }) \\
\leq & \int\limits_{\Omega} (f_{Y|\{d_Y\}} - \mathbf{f}_{Y|\{d_Y\}}) \log\left(f_{Y|\{d_Y\}}\right) d\nu \\
= & \int\limits_{\Omega} (f_{Y|\{d_Y\}} - \mathbf{f}_{Y|\{d_Y\}}) \left(\sum_{k=0}^{m} \lambda_k y^k\right) \nu(dy) \\
= & \sum_{k=0}^{m} \lambda_k \int\limits_{\Omega} y^k \left(f_{Y|\{d_Y\}} - \mathbf{f}_{Y|\{d_Y\}}\right) \nu(dy) \\
= & \sum_{k=0}^{m} \lambda_k \left(\mu_Y^{(k)} - \mu_Y^{(k)}\right) \\
= & 0
\end{aligned}
$$

which clarifies the very fact that $H\left(\mathbf{P}_{Y|\{d_Y\}}\right) \leq H\left(\mathbf{P}^{MEP}_{Y|\{d_Y\}}\right)$ and this proves our **theorem 3.3.1**. □

Appendix B

The role of the Hankel matrix

In this chapter, we shall **reprove** the uniqueness of the solution of the system of equations (4.2) (or equivalently the system of equations (4.4)), **just** for the sake of showing that the positive definiteness of the Hankel matrix proves uniqueness of of the system of equations (4.2).

Before we go ahead, just for the sake of clarity, we restate that the positive definiteness of the Hankel matrix **implies and implied by** the positive definiteness of the matrix $(\sigma_{i,j})_{1\leq i\,,\,j\leq n}$ for $n \in \mathbb{N}$.

B.1 The important lemma

Definition B.1.1. *In general, μ_n ($n \in \mathbb{N}_0$) is defined by*

$$\mu_n = \frac{\int\limits_{\mathcal{X}_X} x^n e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} \tag{B.1}$$

Proposition B.1.1. *With regard to the **definition B.1.1**, for the special case of $n = 1$, β_1 is **uniquely** determinable for a fixedly chosen $\mu_1 = \mu_1^*$ and fixedly chosen $\beta_2 = \beta_2^*, \beta_3 = \beta_3^*, \ldots, \beta_m = \beta_m^*$.*

*Proof of the **proposition B.1.1**.* With subject to $n = 1$,

$$d\mu_1 = \frac{\partial\mu_1}{\partial\beta_1}d\beta_1 + \frac{\partial\mu_1}{\partial\beta_2}d\beta_2 + \ldots + \frac{\partial\mu_1}{\partial\beta_m}d\beta_m \tag{B.2}$$

In that case, for any fixed values of $\beta_2, \beta_3, \ldots, \beta_m$, namely $\beta_2^*, \beta_3^*, \ldots, \beta_m^*$ respectively, thereby implying $d\beta_2 = d\beta_3 = \ldots = d\beta_m = 0$, the above relation (B.2) reduces to

$$d\mu_1 = \frac{\partial \mu_1}{\partial \beta_1} d\beta_1 \tag{B.3}$$

and thus, we get nothing different from

$$\frac{d\mu_1}{d\beta_1} = \frac{\partial \mu_1}{\partial \beta_1} = \mu_2 - \mu_1^2 > 0 \tag{B.4}$$

where in the special case for $\beta_2 = \beta_2^*, \beta_3 = \beta_3^*, \ldots, \beta_m = \beta_m^*$, the moments μ_1 and μ_2 are defined by (B.1) as usual.

This means nothing different from the very fact that for any arbitrarily fixed set of values $\beta_2 = \beta_2^*, \beta_3 = \beta_3^*, \ldots, \beta_m = \beta_m^*$, μ_1 is a strictly monotonically increasing function of β_1, implying that,

If any fixed value of μ_1, say μ_1^*, is taken, then the value of β_1 contained in the right hand side of the expression of μ_1 defined by (B.1) for $n = 1$ can be uniquely determined, provided that the values of $\beta_2, \beta_3, \ldots, \beta_m$ are kept fixed. (B.5)

Conclusively, for a given $\mu_1 = \mu_1^*$, this unique value of β_1, namely β_1^*, is given by

$$\mu_1^* = \frac{\int\limits_{\mathcal{X}_X} x e^{\beta_1^* x + \beta_2^* x^2 + \ldots + \beta_m^* x^m} \nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1^* x + \beta_2^* x^2 + \ldots + \beta_m^* x^m} \nu_X(dx)} \tag{B.6}$$

and this proves the proposition **proposition B.1.1**. □

B.2 Uniqueness of the solution of the equation-system for $m \in \mathbb{N}_0$

Finally, we arrive at the general statement of our theorem, i.e. for every $m \in \mathbb{N}_0$, provided $N \geq m+1$ in discrete cases:

Theorem B.2.1 (Uniqueness of the solution of the simultaneous system of equations of moments). *The solution of the system of m simultaneous equations given in (4.4), namely*

$$\begin{cases} \mu_1 &= \frac{\int\limits_{\mathcal{X}_X} x\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} \\ \mu_2 &= \frac{\int\limits_{\mathcal{X}_X} x^2\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} \\ \vdots & \\ \mu_m &= \frac{\int\limits_{\mathcal{X}_X} x^m\ e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)}{\int\limits_{x\in\mathcal{X}_X} e^{\beta_1 x+\beta_2 x^2+\ldots+\beta_m x^m}\nu_X(dx)} \end{cases} \qquad (4.4)$$

is unique, provided $N \geq m+1$ holds in discrete cases.

Proof of the ***theorem B.2.1****.* By taking the differential of μ_i, $i = 1, 2, \ldots, m$ with respect to $\beta_1, \beta_2, \ldots, \beta_m$, we get

$$d\mu_i = \frac{\partial \mu_i}{\partial \beta_1} d\beta_1 + \frac{\partial \mu_i}{\partial \beta_2} d\beta_2 + \ldots + \frac{\partial \mu_i}{\partial \beta_m} d\beta_m \qquad (B.7)$$

which leads us to the following matrix relation:

$$\begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_1}{\partial \beta_2} & \cdots & \frac{\partial \mu_1}{\partial \beta_m} \\ \frac{\partial \mu_2}{\partial \beta_1} & \frac{\partial \mu_2}{\partial \beta_2} & \cdots & \frac{\partial \mu_2}{\partial \beta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_m}{\partial \beta_1} & \frac{\partial \mu_m}{\partial \beta_2} & \cdots & \frac{\partial \mu_m}{\partial \beta_m} \end{pmatrix} \begin{pmatrix} d\beta_1 \\ d\beta_2 \\ \vdots \\ d\beta_m \end{pmatrix} = \begin{pmatrix} d\mu_1 \\ d\mu_2 \\ \vdots \\ d\mu_m \end{pmatrix} \qquad (B.8)$$

At this point, we state that the proof of the aforesaid uniqueness can be given broadly in three steps:

First Step: In the first step, we shall show that for any fixed values of $\mu_1, \mu_2, \ldots, \mu_m$, say $\mu_1^*, \mu_2^*, \ldots, \mu_m^*$ respectively, the value of β_m can be determined uniquely, say $\beta_m = \beta_m^*$:

For this, let the values of $\mu_1, \mu_2, \ldots, \mu_{m-1}$ be kept fixed, say $\mu_1 = \mu_1^*$, $\mu_2 = \mu_2^*, \ldots, \mu_{m-1} = \mu_{m-1}^*$ and μ_m be allowed to vary.

Therefore, by rewriting the above matrix (B.8) for a given fixed set of values of $\mu_1, \mu_2, \ldots, \mu_{m-1}$, i.e. $d\mu_1 = d\mu_2 = \ldots = d\mu_{m-1} = 0$, we get

$$\begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_1}{\partial \beta_2} & \cdots & \frac{\partial \mu_1}{\partial \beta_{m-1}} & \frac{\partial \mu_1}{\partial \beta_m} \\ \frac{\partial \mu_2}{\partial \beta_1} & \frac{\partial \mu_2}{\partial \beta_2} & \cdots & \frac{\partial \mu_2}{\partial \beta_{m-1}} & \frac{\partial \mu_2}{\partial \beta_m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial \mu_{m-1}}{\partial \beta_1} & \frac{\partial \mu_{m-1}}{\partial \beta_2} & \cdots & \frac{\partial \mu_{m-1}}{\partial \beta_{m-1}} & \frac{\partial \mu_{m-1}}{\partial \beta_m} \\ \frac{\partial \mu_m}{\partial \beta_1} & \frac{\partial \mu_m}{\partial \beta_2} & \cdots & \frac{\partial \mu_m}{\partial \beta_{m-1}} & \frac{\partial \mu_m}{\partial \beta_m} \end{pmatrix} \begin{pmatrix} d\beta_1 \\ d\beta_2 \\ \vdots \\ d\beta_{m-1} \\ d\beta_m \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ d\mu_m \end{pmatrix} \tag{B.9}$$

and thus by using the **Cramer's rule for determinants**, we get

$$d\beta_m = \frac{\det\begin{pmatrix} \frac{\partial\mu_1}{\partial\beta_1} & \frac{\partial\mu_1}{\partial\beta_2} & \cdots & \frac{\partial\mu_1}{\partial\beta_{m-1}} & 0 \\ \frac{\partial\mu_2}{\partial\beta_1} & \frac{\partial\mu_2}{\partial\beta_2} & \cdots & \frac{\partial\mu_2}{\partial\beta_{m-1}} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial\mu_{m-1}}{\partial\beta_1} & \frac{\partial\mu_{m-1}}{\partial\beta_2} & \cdots & \frac{\partial\mu_{m-1}}{\partial\beta_{m-1}} & 0 \\ \frac{\partial\mu_m}{\partial\beta_1} & \frac{\partial\mu_m}{\partial\beta_2} & \cdots & \frac{\partial\mu_m}{\partial\beta_{m-1}} & d\mu_m \end{pmatrix}}{\det\begin{pmatrix} \frac{\partial\mu_1}{\partial\beta_1} & \frac{\partial\mu_1}{\partial\beta_2} & \cdots & \frac{\partial\mu_1}{\partial\beta_{m-1}} & \frac{\partial\mu_1}{\partial\beta_m} \\ \frac{\partial\mu_2}{\partial\beta_1} & \frac{\partial\mu_2}{\partial\beta_2} & \cdots & \frac{\partial\mu_2}{\partial\beta_{m-1}} & \frac{\partial\mu_2}{\partial\beta_m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial\mu_{m-1}}{\partial\beta_1} & \frac{\partial\mu_{m-1}}{\partial\beta_2} & \cdots & \frac{\partial\mu_{m-1}}{\partial\beta_{m-1}} & \frac{\partial\mu_{m-1}}{\partial\beta_m} \\ \frac{\partial\mu_m}{\partial\beta_1} & \frac{\partial\mu_m}{\partial\beta_2} & \cdots & \frac{\partial\mu_m}{\partial\beta_{m-1}} & \frac{\partial\mu_m}{\partial\beta_m} \end{pmatrix}} = \frac{\det(\sigma_{i,j})_{1\leq i\,,\,j\leq m-1}}{\det(\sigma_{i,j})_{1\leq i\,,\,j\leq m}}\,d\mu_m \tag{B.10}$$

which leads us to

$$\frac{d\mu_m}{d\beta_m} = \frac{\det(\sigma_{i,j})_{1\leq i\,,\,j\leq m}}{\det(\sigma_{i,j})_{1\leq i\,,\,j\leq m-1}} > 0 \tag{B.11}$$

simply because we know from our lastly established positive definiteness of the matrix $(\sigma_{i,j})_{1\leq i\,,\,j\leq n}$ for $n \in \mathbb{N}$, but $N \geq n+1$ in case X happens to be discrete, implying $\det(\sigma_{i,j})_{1\leq i\,,\,j\leq n} > 0$ for $n \in \{m-1, m\}$.

Thus, in the language of differential calculus, μ_m is strictly monotonically increasing with respect to the increase in β_m for any given fixed values of $\mu_1, \mu_2, \ldots, \mu_{m-1}$.

This necessarily means, for a given value of μ_m, say $\mu_m = \mu_m^*$, there is an unique value of β_m, say $\beta_m = \beta_m^*$ with subject to the given arbitrarily fixed $\mu_1 = \mu_1^*, \mu_2 = \mu_2^*, \ldots, \mu_{m-1} = \mu_{m-1}^*$ within their individual valid ranges.

In other words, for any chosen fixed values of $\mu_1, \mu_2, \ldots, \mu_m$, namely $\mu_1^*, \mu_2^*, \ldots, \mu_m^*$ respectively, there can exist only an unique value of β_m, namely β_m^*.

Second Step: In the second step, we shall show that for the aforesaid fixed values of $\mu_1, \mu_2, \ldots, \mu_m$, namely $\mu_1^*, \mu_2^*, \ldots, \mu_m^*$ respectively, the values of $\beta_2, \beta_3, \ldots, \beta_{m-1}$ can be determined uniquely, say $\beta_2^*, \beta_3^*, \ldots, \beta_{m-1}^*$:

We have already established that $\beta_m = \beta_m^*$ is a part of the unique solution of (4.4) with subject to the given fixed values $\mu_1^*, \mu_2^*, \ldots, \mu_m^*$. At this point, with regard to the knowledge of the fixed $\mu_m = \mu_m^*$ and $\beta_m = \beta_m^*$, we can actually proceed to work by reducing the system of m simultaneous equations (B.9) to a system of $m-1$ simultaneous equations with respect to the unknowns $\beta_1, \beta_2, \ldots, \beta_{m-1}$, where only $\mu_1, \mu_2, \ldots, \mu_{m-1}$ are contained.

With subject to the consideration of this new system of $m-1$ simultaneous equations, exactly in the same way as before, by considering the fixed values $\mu_1 = \mu_1^*, \mu_2 = \mu_2^*, \ldots, \mu_{m-2} = \mu_{m-2}^*$, but by allowing μ_{m-1} to vary, we get

$$\frac{d\mu_{m-1}}{d\beta_{m-1}} = \frac{\det(\sigma_{i,j})_{1\leq i\,,\,j\leq m-1}}{\det(\sigma_{i,j})_{1\leq i\,,\,j\leq m-2}} > 0 \tag{B.12}$$

after having kept the following under consideration:

- $d\beta_m = 0$, as $\beta_m = \beta_m^*$ is taken for fixed and therefore not considered as a variable in the system of $m-1$ simultaneous equations mentioned immediately above.

- with subject to $i = 1, 2, \ldots, m-1$ the expression of the differential of $d\mu_i$ given in (B.7) is reduced to

$$d\mu_i = \frac{\partial \mu_i}{\partial \beta_1} d\beta_1 + \frac{\partial \mu_i}{\partial \beta_2} d\beta_2 + \ldots + \frac{\partial \mu_i}{\partial \beta_{m-1}} d\beta_{m-1} \tag{B.13}$$

- for $\mu_1, \mu_2, \ldots, \mu_{m-2}$ being fixed, $d\mu_1 = d\mu_2 = \ldots d\mu_{m-2} = 0$

- the very proven assertion stating that $\det(\sigma_{i,j})_{1\leq i\,,\,j\leq n} > 0$ for $n \in \mathbb{N}$, but $N \geq n+1$ for $n \in \{1, 2, \ldots, m-1\}$ in case X happens to be discrete

which again leads us to conclude that μ_{m-1} is strictly monotonically increasing with respect to the increase in β_{m-1} for the given fixed values of $\mu_1, \mu_2, \ldots, \mu_{m-2}$.

This again necessarily means, for a given value of μ_{m-1}, say $\mu_{m-1} = \mu^*_{m-1}$, there is an unique value of β_{m-1}, say $\beta_{m-1} = \beta^*_{m-1}$ with subject to the given fixed $\mu_1 = \mu^*_1, \mu_2 = \mu^*_2, \ldots, \mu_{m-2} = \mu^*_{m-2}$ within their individual valid ranges.

In other words, for the aforesaid chosen fixed values of $\mu_1, \mu_2, \ldots, \mu_{m-1}$, namely $\mu^*_1, \mu^*_2, \ldots, \mu^*_{m-1}$ respectively, there can exist only an unique value of β_{m-1}, namely β^*_{m-1}.

Thus, we have additionally established that $\beta_{m-1} = \beta^*_{m-1}$ is too a part of the unique solution of (4.4).

Proceeding exactly in this manner as above, we arrive at

For the aforesaid chosen fixed values of μ_1, μ_2, namely μ^*_1 and μ^*_2 respectively, there can exist only an unique value of β_2, namely β^*_2.

Therefore, till this point, we have established the very fact that, with subject to $\mu_1 = \mu^*_1, \mu_2 = \mu^*_2, \ldots, \mu_m = \mu^*_m$, the values of $\beta_2, \beta_3, \ldots, \beta_m$ can be uniquely determined, namely $\beta^*_2, \beta^*_3, \ldots, \beta^*_m$ respectively, thereby leading us to conclude that $\{\beta_2 = \beta^*_2, \beta_3 = \beta^*_3, \ldots, \beta_m = \beta^*_m\}$ is too a part of the unique solution of (4.4).

Third Step: In the third and the final step, we shall show that the value of β_1 contained in the system (4.4) is unique too, namely β^*_1. Immediately after this, the targeted uniqueness follows conclusively.

Here, by using (B.5) (of the **proposition B.1.1**), for the uniquely determined $\beta_2 = \beta^*_2, \beta_3 = \beta^*_3, \ldots, \beta_m = \beta^*_m$ we can undoubtedly conclude that the value of β_1 is unique, which satisfies

$$\mu^*_1 = \frac{\int\limits_{\mathcal{X}_X} x e^{\beta_1 x + \beta^*_2 x^2 + \ldots + \beta^*_m x^m} \nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1 x + \beta^*_2 x^2 + \ldots + \beta^*_m x^m} \nu_X(dx)} \tag{B.14}$$

for say $\beta_1 = \beta^*_1$.

This is nothing, but to say that

$$\mu^*_i = \frac{\int\limits_{\mathcal{X}_X} x^i e^{\beta^*_1 x + \beta^*_2 x^2 + \ldots + \beta^*_m x^m} \nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta^*_1 x + \beta^*_2 x^2 + \ldots + \beta^*_m x^m} \nu_X(dx)} \text{ , for } i \in \{2, 3, \ldots, m\} \tag{B.15}$$

where β_1^* is uniquely determined by

$$\mu_1^* = \frac{\int\limits_{\mathcal{X}_X} x e^{\beta_1^* x + \beta_2^* x^2 + \ldots + \beta_m^* x^m} \nu_X(dx)}{\int\limits_{\mathcal{X}_X} e^{\beta_1^* x + \beta_2^* x^2 + \ldots + \beta_m^* x^m} \nu_X(dx)} \tag{B.16}$$

Hence, in other words, for given values of $\mu_1, \mu_2, \ldots, \mu_m$, namely $\mu_1^*, \mu_2^*, \ldots, \mu_m^*$ respectively, $\{\beta_1 = \beta_1^*, \beta_2 = \beta_2^*, \ldots, \beta_m = \beta_m^*\}$ is the only solution of the system (4.4).

Whence, the uniqueness of the solution of (4.4) has been proved and thus the **theorem B.2.1** gets duly proved. □

Importantly, this uniqueness of the solution of (4.4) is tantamount to the uniqueness of the solution of the system of equations (4.2).

Remark B.2.1 (Concluding remark). *Therefore, it is conclusively clear that the uniqueness of the solution of the system of m equations (4.4) or the uniqueness of the solution of the system of m equations (4.2), is basically proved by the positive definiteness of the symmetric matrix* $(\sigma_{i,j})_{1 \leq i\,,\, j \leq m}$.

Appendix C

Miscellaneous

C.1 Inequalities

Theorem C.1.1 (Holder's inequality). *If* $\frac{1}{p} + \frac{1}{q} = 1$ *for* p *and* q *being two positive real numbers and if* $f(x)$ *and* $g(x)$ *are two real valued functions, then*

$$\left| \int_a^b f(x)\, g(x)\, dx \right| \leq \left\{ \int_a^b |f(x)|^p\, dx \right\}^{\frac{1}{p}} \left\{ \int_a^b |g(x)|^q\, dx \right\}^{\frac{1}{q}}$$

Referred to the page 139 of [34].

Remark C.1.1. *Understandably,* $1 \leq p \leq \infty$ *and* $1 \leq q \leq \infty$ *must hold, after having* ***additionally*** *defined the following:*

$$\left\{ \int_a^b |f(x)|^r\, dx \right\}^{\frac{1}{r}} = \sup_{x \in [a,b]} |f(x)| \ \textit{ in case } r = \infty$$

Corollary C.1.1 (Cauchy-Schwarz inequality). *In the* ***special case*** *of* $p = q = 2$*, the Holder's inequality becomes the Cauchy-Schwarz inequality, which is given as follows:*

$$\left| \int_a^b f(x)\, g(x)\, dx \right| \leq \left\{ \int_a^b |f(x)|^2\, dx \right\}^{\frac{1}{2}} \left\{ \int_a^b |g(x)|^2\, dx \right\}^{\frac{1}{2}}$$

Bibliography

[1] Apostol, Tom M. (1969): *Calculus, Volume II*. Blaisdell Publishing Company; Second Edition · Massachusetts · Toronto · London

[2] Ash, Robert B. (1965): *Information Theory*, Interscience Publ. New York

[3] Balestrino, A., Caiti, A., Noe, A., Parenti, F. (2003): *A Class of Numerical Algorithms for Efficient Approximation of Maximum Entropy Estimates of Probability Density Functions*. 11 th Mediterranean Conference on Control and Automation.

[4] Bandyopadhyay, K., Bhattacharya, A. K., Biswas, P. and Drabold, D. A. (2005): *Maximum entropy and the problem of moments: A stable algorithm*. Physical Review E., Vol. 71, pp 057701-1 ... 057701-4.

[5] Barner Martin, Flohr Friedrich (1982): Analysis II. de Gruyter Lehrbuch Verlag. Walter de Gruyter, Berlin, New York, 148-153.

[6] Bartoszyński, Robert and Niewiadomska-Bugaj, Magdalena (1996): *Probability and Statistical Inference*. John Wiley & Sons, Inc. New York, Chichester, Brisbane, Toronto, Singapore.

[7] Beck, Andreas (1989/90): *Skriptum zur Vorlesung Informatik III von Prof. Dr. Ing. P. Tran-Gia*

[8] Bityutskov, V. I. (2001): *Bunyakovskii inequality*, in Hazewinkel, Michiel, *Encyclopaedia of Mathematics*, Kluwer Academic Publishers, ISBN 978-1556080104

[9] Borwein, J. M. & Lewis, A. S. (1991): *Convergence of best entropy estimates*. SIAM J. Optimization 1: 191 - 205

[10] Bundesverband WindEnergie e.V. [www.wind-energie.de]

[11] Cover, Thomas M. and Thomas, Joy A. (2006): *Elements of Information Theory*. Second Edition. Published by John Wiley & Sons, Inc., Hoboken, New Jersey. Published in Canada.

[12] Goeb, Rainer (2009): *Matrizentheorie*

[13] Gupta, Amritava (1983): *Groundwork of Mathematical Probability and Statistics*. Academic Publishers, Calcutta, New Delhi.

[14] Frontini, Marco & Tagliani, Aldo: *Entropy- Convergence in Stieltjes and Hamburger moment problem*. Dipartimento di Mathematica, Politecnico di Milano, Piazza L. da Vinci, 32 20133 Milano, Italy.

[15] Frontini, Marco & Tagliani, Aldo (1994): *Maximum Entropy in the finite Stieltjes and Hamburger moment problem*. Dipartimento di Mathematica, Politecnico di Milano, Piazza L. da Vinci, 32, 23100 Milano, Italy.

[16] Ghosh, Ramkrishna and Maity, Kantish Chandra (1969): *Differential Calculus, An Introduction to Analysis*. New central book agency.

[17] Ghosh, Ramkrishna and Maity, Kantish Chandra (1981): *Integral Calculus, An Introduction to Analysis*. New central book agency.

[18] Google search "Hamburger Moment Problem - Wikipedia"

[19] Hausdorff, Felix (1921): Summationsmethoden und Momentfolgen I, II. *Mathematische Zeitschrift* 9 Band, 74-109, 280-299. Verlag vom Julius Springer, Berlin

[20] Karlin, S. and Shapley L. S. (1953): *Geometry of moment spaces*, Mem. Amer. Math. Soc., 12

[21] Hilbert, David (1900): Mathematical Problems. Lecture delivered before the International Congress of Mathematicians at Paris in 1900

[22] Hilbert, David (1999): *Grundlagen der Geometrie*. B. G. Teubner Stuttgart · Leipzig 1999, 303 - 315, 364 - 368.

[23] Hofstadter, Douglas R. (1985): *Gödel, Escher, Bach: Ein Endloses Geflochtenes Band*. Klett - Cotta Verlag, Stuttgart 1985, 96 - 103.

(The original English edition was published under the title *An Eternal Golden Braid*. Gödel, Escher, Bach ©1979 by Basic Books, New York.)

[24] Jaynes, E.T. (1957): *Information Theory and Statistical Mechanics. Phys. Rev.* 106, 620-630; 108, 171-182.

[25] Kapur, J.N. (1993): *Maximum-Entropy Models in Science and Engineering*. Wiley Eastern Limited.

[26] Kolmogorov, Andrey Nikolaevich (1956): *Foundations of the Theory of Probability, Second English Edition*. Chelsa Publishing Company, New York.

[27] Loney, S. L. (1981): *An Elementary Treatise on the Dynamics of a particle and of rigid bodies* 10-12. Macmillan India Limited

[28] Mitra A. (2004): Mathematics and its Foundations. PhD thesis. Arcata, California. Appeared in the article "Journey in Being".

[29] Naas J. und Schmid H.L. (1965): *Mathematisches Wörterbuch, Band 2*, 193 - 195. Akademie -Verlag GmbH - Berlin & B. G. Teubner Verlagsgesellschaft - Stuttgart

[30] Pieper, John: *Entropy, Disorder and Life*. Copyright © 2000.
Updated: May 24, 2002

[31] Chen, B., Hu, J. and Zhu, Y. (2010): *Computing Maximum Entropy Densities: A Hybrid Approach* Signal Processing: An Internationl Journal (SPIJ), Volume (4); Issue (2), pp 114-122.

[32] Jeffreys, Sir Harold and Swirles, Bertha (Lady Jeffreys) (1988) "**Weierstrass's theorem on approximation by polynomials**" and "**Extension of Weierstrass's approximation theorem**" §14.08 - 14.081: *Methods of Mathematical Physics* 3rd Edition Cambridge, England: Cambridge University Press, pp 446 - 448.

[33] Renyi, Alfred (1970): *Probability Theory* Volume 10, North-Holland publishing company - Amsterdam . London, pp 540 - 605

[34] Rudin, Walter (1976): *Principles of mathematical analysis* Professor of mathematics, University of Wisconsin - Madison, third edition, ISBN 0-07-054235-X. Copyright ©1976

[35] Sastry, S. S. (1983): *Introductory methods of numerical analysis* Prentice-Hall of India Private Limited, New Delhi-110001

[36] Scarborough, James Blaine (1930): *Numerical Mathematical Analysis* 17, 215-219. Oxford and IBH Publishing Co., New Delhi, Bombay, Calcutta

[37] Sen, Surath (2001): *Diplomarbeit, Surath Sen* Universität Würzburg

[38] The magister programme of the e-learning system of the company Stochastikon GmbH, Würzburg, Germany conducted by: http://www.stochastikon.com

[39] Surath Sen and Elart von Collani (2007): A Note on the Moments of Random Variables. *Economic Quality Control* 22, Number 2, 223 - 246

[40] Stoer, Josef (1989): *Numerische Mathematik 1* 5th edition, Springer Verlag, Berlin Heidelberg, New York, London, Paris, Tokyo, HongKong, pp 114 - 261

[41] Tagliani, A. (1999): *Hausdorff Moment Problem and Maximum Entropy: A unified Approach.* Facoltà di Economia Università di Trento, 38160 Trento, Italy · Applied Mathematics and Computation 105 (1999) 291 - 305

[42] von Collani, E., Binder, Sans, Heltmann, Al-Ghazali, Böhme: *Design Load Definition for Wind Turbines by LEXPOL* in Wind Energy (Special Issue: Design Load Definition), Wiley Interscience. eingereicht 2007 (WE-07-0087)

[43] von Collani, E. (2004): Theoretical Stochastics. In *Defining the Science of Stochastics.* Ed. by E. v. Collani. Heldermann Verlag, Lemgo, 147-174.

[44] von Collani, E. (2004): Empirical Stochastics. In *Defining the Science of Stochastics.* Ed. by E. v. Collani. Heldermann Verlag, Lemgo, 175-213.

[45] von Collani, E. (2004): History, State of the Art and Future of Stochastics. In *History of the Mathematical Sciences.* Ed. by I. Grattan-Guinness B.S. Yadav. Hindustan Book Agency, 171-194.

[46] von Collani, E. and Dumitrescu, M. (2000): Neyman Exclusion Procedure, *Economic Quality Control*, 15, 15-34.

[47] von Collani, E. and Dumitrescu, M. (2000): Neyman Comparison Procedure, *Economic Quality Control*, 15, 35-53.

[48] von Collani, E., Dumitrescu, M. and Lepenis, R. (2001): Neyman Measurement and Prediction Procedures, *Economic Quality Control*, 16, 109-132.

[49] von Collani, E. and Dumitrescu, M. (2001): Complete Neyman Measurement Procedure, *Metrika*, 54, pp.111-130.

[50] von Collani and E., Dräger, K. (2001): *Binomial Distribution. Handbook for Scientists and Engineers*, Birkhäuser, Boston.

[51] von Collani, E., Dumitrescu, M. and Panaite, V. (2002): Prediction and Measurement Procedures for the Variance of a Normal Distribution, *Economic Quality Control*, 17, 133-154.

[52] von Collani, E. and Binder, A. (2002): Einführung in die Theoretische Stochastik. Würzburg Research Group on Quality Control (WRQC), pp. 13.

[53] von Collani, E. and Binder, A. (2002): Einführung in die Empirische Stochastik - Teil 1. Würzburg Research Group on Quality Control (WRQC)

[54] edited by von Collani, E. (2003): Theoretical Stochastics. In the book titled by *Defining the Science of Stochastics*. Sigma Series in Stochastics Volume 1, 147-174. Copyright Heldermann Verlag.

[55] edited by von Collani, E. (2003): Empirical Stochastics. In the book titled by *Defining the Science of Stochastics*. Sigma Series in Stochastics Volume 1, 175-213. Copyright Heldermann Verlag.

[56] edited by von Collani, E. (2003): *Defining the Science of Stochastics*. Sigma Series in Stochastics Volume 1. Copyright Heldermann Verlag.

[57] Wallace, Edward & West, Stephen (2003): *Roads to Geometry*. Prentice Hall; Third Edition, Appendix B.

[58] Weierstrass, Karl (1885): Über die analytische Darstellbarkeit sogenannter willkürlicher Funktionen einer reellen Veränderlichen. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, 1885 (II).

[59] Ximing, Wu (2003): *Calculation of maximum entropy densities with application to income distribution*. Department of Agricultural and Resource Economics, University of California at Berkeley, Berkeley, CA 94720, USA, and Department of Economics, University of Guelph, Ont., Canada. Journal of Econometrics 115 (2003) 347 - 354.

[60] Young Hwa Sung and Byung Man Kwak (2010): *Reliability bound based on the maximum entropy principle with respect to the first truncated moment* Department of Mechanical Engineering, KAIST, Daejeon, 305 - 701, Korea.