Jahresbericht der Deutschen Mathematiker-Vereinigung (2022) 124:137–143
https://doi.org/10.1365/s13291-022-00250-y

BOOK REVIEW

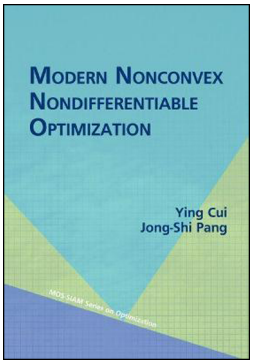
Y. Cui, J.-S. Pang: "Modern Nonconvex Nondifferentiable Optimization"

SIAM, 2022, xx + 756 pp.

Christian Kanzow[1]

Accepted: 28 March 2022 / Published online: 4 April 2022

1 Motivation



Before discussing the contents of the book under review, let me start with some motivational remarks and illustrative examples in order to get a feeling of the difficulties that arise in the context of nonconvex and nondifferentiable optimization problems. In principle, one distinguishes between an unconstrained minimization problem

$$\min\ f(x), \quad x \in \mathbb{R}^n, \tag{1}$$

for some given objective function $f : \mathbb{R}^n \to \mathbb{R}$, and a constrained optimization problem

$$\min\ f(x) \quad \text{subject to} \quad x \in X, \tag{2}$$

where the feasible set X is a subset of $\mathbb{R}^n$. Both problems are supposed to be easy ones if they are convex, which means that f is convex and, for problem (2), the feasible set is also convex. Convexity makes these problems simple since all local minima are automatically global minima. In fact, even every stationary point can be shown to be a global minimum, and there is a whole bunch of algorithms which are able to find stationary points of convex optimization problems.

On the other hand, this statement already yields a first question: What is a stationary point of, say, the unconstrained problem (1) if f is nonsmooth? For smooth objective functions, x^* is called a stationary point if it satisfies the standard optimality condition $\nabla f(x^*) = 0$. For nonsmooth f, on the other hand, one has to use other conditions. One possibility would be to require that $f'(x^*; d) \geq 0$ for all d, i.e.,

✉ C. Kanzow
kanzow@mathematik.uni-wuerzburg.de

[1] Würzburg, Germany

the directional derivative needs to be nonnegative in all directions, provided that this directional derivative exists. Alternatively, one can use the convex subdifferential

$$\partial f(x^*) := \{s \mid f(x) \geq f(x^*) + s^T(x - x^*) \ \ \forall x \in \mathbb{R}^n\}$$

which, geometrically, consists of all slopes s such that the affine function on the right-hand side is below the mapping f and touches this function at the point x^*. Then $0 \in \partial f(x^*)$ is another natural generalization of stationarity for convex functions which, in this particular case, coincides with the previous one based on the directional derivative. In a similar way, it is possible to derive stationarity conditions for the constrained problem (2).

Now, what happens in the nonconvex case? First of all, nonconvex problems usually have local minima which are not global ones. Moreover, a stationary condition based on the directional derivative can be directly extended to general nonconvex objective functions and also yields a necessary optimality condition there, provided the directional derivative exists. Furthermore, one can also generalize the convex subdifferential to nonconvex functions, but it turns out that this is a delicate problem since there exist several generalizations having different properties. To illustrate the difficulty, let us define a very simple one, say $\partial f(x) := \mathbb{R}^n$ for all x. This simple subdifferential has very nice properties, it is always nonempty, the sum rule $\partial(f+g)(x) = \partial f(x) + \partial g(x)$ as well as several other calculus rules hold. In addition, if x^* denotes a local minimum, then the necessary condition $0 \in \partial f(x^*)$ is satisfied. Nevertheless, this subdifferential is completely useless since every point $x \in \mathbb{R}^n$ satisfies the necessary optimality condition $0 \in \partial f(x)$, i.e., this subdifferential does not distinguish between good and bad candidates for a local minimum. Consequently, the main idea is to find a subdifferential which is as small as possible, but still possesses at least some calculus rules.

We next present two illustrative examples which should explain some further ideas. The first one is the portfolio problem by Markowitz

$$\min \frac{1}{2} x^T Q x \quad \text{subject to} \quad \mu^T x \geq \rho, \ e^T x = 1, \ x \geq 0,$$

where e is the all-one vector, Q and μ denote the covariance matrix and the mean of n possible assets/stocks, respectively, while ρ is some lower bound for the expected return. Hence the objective is, basically, to minimize the risk, while having a guaranteed expected return if we investigate a certain part (given by x_i) of our total money (which sums up to 1) into asset i. Note that Q, being a covariance matrix, is always positive semidefinite, hence the portfolio optimization problem is a convex constrained minimization problem. Now, due to some additional side costs etc, one typically invests the money such that only a few components x_i are really positive, i.e., one is looking for a sparse solution of the portfolio problem. Usually, this is modeled by adding the ℓ_1-norm to the objective function, thus we obtain the modified problem

$$\min \frac{1}{2} x^T Q x + \alpha \|x\|_1 \quad \text{subject to} \quad \mu^T x \geq \rho, \ e^T x = 1, \ x \geq 0$$

for some weight $\alpha > 0$, since it is known that adding such a term (usually) improves the sparsity of the solution. Moreover, adding the ℓ_1-norm keeps the convexity of the portfolio problem, though it becomes nondifferentiable. But due to the particular constraints, this sparsity term is just a constant in our case. This motivates to replace the ℓ_1-norm either by the ℓ_0- or the ℓ_p-quasi-norm for $p \in (0, 1)$, where $\|x\|_0$ counts the number of nonzero components of x, whereas $\|x\|_p$ is defined in the usual way, except that $p < 1$. The corresponding formulation can indeed be used to yield sparse solutions, but the resulting problems are now nonconvex and nondifferentiable (in case of the ℓ_0-term even discontinuous), i.e., these problems are much more realistic models to get sparse solutions, but they are also significantly more difficult to solve.

As a second example, consider the (unconstrained) DC program (DC = difference-of-convex)

$$\min \ f(x) - g(x), \quad x \in \mathbb{R}^n,$$

where the objective function is the difference of two convex mappings f and g and, hence, nonconvex (and possibly nondifferentiable). The corresponding DC algorithm is an iterative method which replaces the single DC program by a sequence of problems of the form

$$\min \ f(x) - g_k(x), \quad x \in \mathbb{R}^n, \tag{3}$$

where $g_k(x) := g(x^k) + (s^k)^T(x - x^k)$ with an arbitrary $s^k \in \partial g(x^k)$ is the lower bound of g defined by the convex subdifferential at the current iterate x^k. Since g_k is an affine function, the objective function of (3) is convex and, therefore, easy to minimize. Using the fact that $f - g \leq f - g_k$, the minimizer x^{k+1} of the majorization function $f - g_k$ might give a good candidate for a minimum of the original DC program. A generalization of this technique is the main idea of the surrogation methods to be described in this book.

2 Contents of the Book

The book considers nonsmooth and nondifferentiable optimization problems, presents the necessary theoretical background and a class of methods for the solution of these problems. Recall, however, that nonconvex problems might have many local minima, and the aim of this book is not to present methods which are able to compute a global minimum. This would be impossible (at least within an acceptable time and without having a very special structure). Suitable methods in continuous optimization are typcially able to compute stationary points and, hence, candidates for a local minimum.

The book is divided into eleven chapters. Chapter 1 presents some general background material from calculus and linear algebra including some basics from a course on numerical mathematics. This material is probably known by most potential readers. In addition, there is a short section on some elements from set-valued analysis, which is less standard and central for some of the subsequent chapters.

Chapter 2 covers some background material from optimization. In particular, some basic theory regarding convex functions and projections are recalled here, as well as the famous KKT optimality conditions (these are generalizations of the Lagrange multiplier rule) and (some) corresponding constraint qualifications for smooth nonlinear programs. In addition, the authors also present a complete convergence theory for the proximal gradient method and a realization of the proximal Newton idea. These are standard methods for solving optimization problems where the objective function is a sum of a smooth and a nonsmooth term, where the emphasis in this background chapter is on convex problems.

Chapter 3, entitled "Structured Learning via Statistics and Optimization", mainly serves as a motivation to deal with nonsmooth and nonconvex optimization problems. It covers a host of applications arising in statistics and optimization. This includes sparse and low rank matrix problems like they occur, for example, in the famous netflix or matrix completion problem, see also the sparse formulation of the portfolio problem. This chapter does not contain any theoretical results.

Since this monograph deals with nonsmooth optimization, it requires some background from nonsmooth analysis. The corresponding material is presented in Chapter 4, which is more than 100 pages long. It starts with different generalizations of the notion of a differentiable function and then presents some particularly relevant classes that will be exploited in the subsequent chapters (like piecewise smooth functions, weakly convex functions, DC-functions, semismooth functions). The corresponding discussion on different generalizations of the notion of a differentiable function to some nonsmooth mappings concentrates, however, on the class of B-differentiable (i.e., directionally differentiable and locally Lipschitz continuous) functions. This is still a very general class of nonsmooth functions, but excludes, for example, a direct treatment of rank minimization problems or optimization problems involving the ℓ_0-quasi-norm (see the introductory example). On the other hand, these classes of problems can often be reformulated in a suitable way so that they still fit within the framework of this monograph.

Chapter 5 is devoted to a discussion of value functions. These value functions arise in several areas in a, more or less, natural way. For example, the (Lagrange) dual of a nonlinear program involves a value function in its objective (which is nonsmooth, but convex in this particular case), gap functions are used to solve variational inequalities, the Nikaido-Isoda function is a value function for Nash equilibrium problems (see also Chapter 11), and a standard approach for investigating and solving bilevel programs is also based on value functions. Robust optimization, dealing with the problem of how robust a solution is subject to suitable perturbations of the given data, is another application area. The corresponding value functions are, in general, nonsmooth and nonconvex, but still have certain smoothness properties, at least under suitable assumptions.

Chapter 6 (almost 170 pages long) considers the highly important topic of stationary points of a given problem. The importance of having suitable notions of stationarity for nonsmooth and nonconvex problems, both unconstrained and constrained ones, was already pointed out in the motivational discussion at the beginning of this report. Despite having these different notions of stationarity, a central practical question is also, which kind of stationarity one can guarantee to obtain by a limit point

of a suitable algorithm. The directional derivative-based stationarity concepts are relatively strong ones, but the mapping $x \mapsto f'(x;d)$ is, in general, nonsmooth, even discontinuous, and therefore causes some difficulties in proving suitable convergence results where one wants to show that a limit point x^* of a sequence $\{x^k\}$ satisfies such a stationarity condition. These considerations hopefully motivate the extensive treatment of this topic in Chapter 6.

Chapter 7 on "Computational Algorithms by Surrogation" is the main algorithmic chapter for the solution of nonsmooth and nonconvex optimization problems. It describes and investigates several surrogation methods for the solution of different kinds of problems. The basic idea is the following: Given the (for simplicity of presentation) unconstrained problem of minimizing a possibly complicated function f, it replaces this single minimization problem by a sequence of minimization problems, where, in each step k, a function f_k needs to be minimized, where f_k is an upper bound of f, i.e., $f \leq f_k$ for all $k \in \mathbb{N}$, and the minimization of f_k itself is easier than the one of f (this is sometimes also called an MM method, with MM standing for the minimization of a majorization function). A particular instance of this idea is the DC algorithm mentioned in the beginning. The surrogation methods extend the idea of the DC algorithm, and particular instances of surrogation functions depend on the structure and (smoothness) properties of the given optimization problem.

Chapter 8 investigates several error bounds whose aim is to provide a computationally inexpensive evaluation of a function which gives an upper bound of the distance of a given point to a certain set, say, the feasible set or the set of solutions/stationary points of a given optimization problem. The most famous result in this area is probably the Hoffman error bound for a polyhedral set, whereas this chapter concentrates on nonpolyhedral and nonconvex problems. A very popular and helpful error bound in this area is given by the Kurdyka-Łojasiewicz theory, which is therefore also covered in this chapter. Error bounds have at least three important applications: They can be used to prove convergence of the entire sequence generated by suitable methods, they imply rate-of-convergence results, and they provide exact penalty results for constrained optimization problems (see the next chapter). The theory of error bounds is closely linked to some regularity notions (linear regularity and metric subregularity), and a corresponding section is therefore also devoted to such a discussion.

The following Chapter 9 presents exact penalty results for constrained optimization problems. Hence, suppose we want to minimize an objective function f on a feasible set given by some set X. Then, most penalty functions p_α are of the form $p_\alpha(x) = f(x) + \alpha r(x)$ for some penalty parameter $\alpha > 0$ and a residual function r which is nonnegative on the whole space and equal to zero exactly on X. Hence the residual function r measures (and, therefore, penalizes) the violation of the constraints. Most text books do not provide a formal definition of exactness, but implicitly call a penalty function exact if a minimum (or stationary point) of the given optimization problem is also a minimum (or a stationary point) of the penalty function p_α for some *finite* penalty parameter $\alpha > 0$ (the converse directions are usually more difficult to prove and require stronger assumptions). In the nonconvex setting, the result on stationary points is obviously much more interesting than a statement on (global) minima, since the former are likely to be computable, whereas the latter are not. But then, of course, the notion of stationarity depends again on the (smoothness)

properties of the underlying program. Basically, this chapter derives several exactness results for different classes of nonconvex (and also some nonsmooth) optimization problems. Since the exact penalty functions themselves are nonsmooth and (usually) nonconvex, also a surrogate-type method is presented for the direct minimization of exact penalty functions.

The remaining two Chapters 10 and 11 present applications of the theory and algorithms to nonconvex stochastic programs and nonconvex Nash equilibrium problems, respectively. The former, in particular, contains optimization problems defined by expectation functions (both in the objective function and the constraints) as well as relatively general two-stage problems, whereas the latter presents some methods for the solution of noncooperative games, where the standard solution concept of a (generalized) Nash equilibrium is replaced by corresponding stationary conditions, in order to deal with the underlying nonconvexity (in particular, depending on the stationary concept used in this context, this leads to the notion of a quasi-Nash equilibrium).

3 Summary

The book gives a comprehensive treatment of nonsmooth and nonconvex optimization problems, and is, to the best of my knowledge, the first monograph exclusively dealing with this kind of problem. Researchers already working in this area or being interested in this subject will benefit a lot from this book. It presents the material in a unified way which, otherwise, can only be found in several (mainly very recent) journal articles or technical reports.

The topics covered in this monograph are, to some extend, dominated or motivated by the idea of solving these difficult optimization problems by surrogation methods. These methods are indeed one of the main and most successful tools for solving nonsmooth and nonconvex minimization problems, and the authors are world-leading experts in this area (well, Jong-Shi Pang, being the more senior author of this book, has also contributed a lot to many other areas in optimization and its applications).

This leads to the only (and very minor) criticism: The book does not cover everything .... Specifically, the class of surrogation methods is one (and important) attempt to solve difficult classes of optimization problems, but not the only one. Proximal-type methods and, in particular, augmented Lagrangian-type methods are nowadays also able to deal with several classes of nonsmooth and nonconvex optimization problems, where the difficulties either arise in the objective function being nonsmooth or the constraints being complicated, sometimes even under weaker smoothness assumptions than B-differentiability. A corresponding treatment, however, would require some notions on the limiting normal cone by Boris Mordukhovich and related results from variational analysis.

On the other hand, the current book is already "heavy" with more than 750 pages ....

Funding Note Open Access funding enabled and organized by Projekt DEAL.