Mathematical Programming: State of the Art 1994 (J.R. Birge and K.G. Murty, editors), University of Michigan Press, Ann Arbor, 1994, pp. 248–258.

NONSMOOTH OPTIMIZATION

Terry Rockafellar, University of Washington

A function is *smooth* if it is differentiable and the derivatives are continuous. More specifically, this is first-order smoothness. Second-order smoothness means that second-derivatives exist and are continuous, and so forth, while infinite smoothness refers to continuous derivatives of all orders. From this perspective a *nonsmooth* function only has a negative description—it lacks some degree of properties traditionally relied upon in analysis. One could get the impression that "nonsmooth optimization" is a subject dedicated to overcoming handicaps which have to be faced in miscellaneous circumstances where mathematical structure might be poorer than what one would like. But this is far from right.

Instead, nonsmooth optimization typically deals with highly structured problems, but problems which arise differently, or are modeled or cast differently, from ones for which many of the mainline numerical methods, involving gradient vectors and Hessian matrices, have been designed. The nonsmoothness can be primary, in the sense of resulting from something deep in the nature of the application at hand, or secondary through the introduction of penalty expressions or various technical subproblems. Anyway, a strong argument can be made for the notion that nonsmoothness in optimization is very often a question of modeling, and due to the prevalence of inequality constraints, is present anyway in almost all problems of importance, at least in the background. The issue from that angle is simply how to make use of available structure in the best possible way. Nonsmooth optimization gives the message that many effective approaches are possible, and one need not be confined to a classical view of how functions are to be approximated and evaluated.

Because nonsmoothness has different manifestations and treatments, one shouldn't imagine that numerical techniques in nonsmooth optimization can act as "black boxes." Techniques are developed for the particular structures that compensate for the absence of differentiability. It's important therefore to understand the source of any nonsmoothness, before deciding how it might be handled. Providing an overview of this issue is one of the main goals in these notes, along with painting a broad picture of the applications and computational ideas characteristic of nonsmooth optimization.

The central fact is that when functions are *defined* in terms of operations of maximization or minimization, in contrast to long-familiar operations of calculus like composition, addition and integration, they may well fail to inherit the smoothness enjoyed by the functions or mappings serving as "ingredients" in the definition. The theory of nonsmooth optimization is largely concerned therefore with extensions of calculus to cover such functions, for instance in terms of generalized directional derivatives and subgradients, and approximation methodology that can be substituted for nonexistent Taylor expansions of first or second order.

Functions with an envelope representation. One of the most common situations is that of minimizing a function f having a representation of the form

$$f(x) = \max_{s \in S} \phi_s(x) \ \text{ for } \ x \in R^n, \tag{1}$$

where S is some index set, finite or infinite—perhaps a subset of R^d as a parameter space—and $\phi_s(x)$ is smooth with respect to x. The likely nonsmoothness of f in such circumstances can be addressed in more than one way.

When S is finite, the minimization of f over a subset C of R^n can be approached in principle by reformulating the given problem in a higher dimensional space. From $x = (x_1, \ldots, x_n) \in R^n$ and an additional variable $x_0 \in R$, one can put together vectors $\tilde{x} = (x_0, x_1, \ldots, x_n) \in R^{n+1}$ and look instead to minimizing $f_0(\tilde{x}) = x_0$ over all $\tilde{x} \in \tilde{C} = R \times C$ that satisfy the constraints

$$f_s(\tilde{x}) \leq 0 \ \text{ for each } \ s \in S, \ \text{ where } \ f_s(\tilde{x}) = -x_0 + \phi_s(x_1, \ldots, x_n). \tag{2}$$

Here f_0 and the constraint functions f_s are smooth, so the problem has been placed in a standard setting. Additional constraints, only in the variables $x_1, \ldots, x_n$, may of course express the set C.

Although this is possible for any finite index set S, the question is whether it is the best way to proceed, and the answer hinges on the size of S as well as the viability of techniques for minimizing f directly. Clearly, when S is very large the proposed reformulation is no panacea. A huge number of constraints, especially nonlinear constraints, isn't easy to cope with. The idea becomes attractive of working instead with subproblems in which a convenient local approximation to f, generated somehow from the envelope representation (1), is minimized over C.

When S is infinite, of course, the reformulation leads to an infinite constraint system and a problem of the kind known as *semi-infinite programming*. Indeed, semi-infinite programming could well be classified as the branch of nonsmooth optimization in which this tactic is applied to an objective function, or possibly an inequality constraint function, having an envelope representation.

The drawback to converting a problem with infinite or very large S to semi-infinite programming, or almost semi-infinite programming, is not only that dual dimensionality is

increased, but that the focus is shifted away from properties of f that might otherwise be put to very good use. This is where ideas for generating approximations to f get interesting. For an introduction to direct numerical methods in this subject (about which we'll have more to say later), the books of Kiwiel [1], Shor [2], and Hiriart-Urruty/Lemarechal [3] are suggested together with the paper of Zowe [4].

Eigenvalue functions. Consider an $m \times m$ symmetric matrix $A(x)$ with entries depending smoothly on $x = (x_1, \ldots, x_n)$ as parameter vector, and let $\lambda_1(x) \geq \lambda_2(x) \geq \cdots \geq \lambda_m(x)$ be the associated eigenvalues (where multiple eigenvalues are repeated). Many applications of optimization involve minimizing a function

$$f(x) = g\big(\lambda_1(x), \cdots, \lambda_m(x)\big), \tag{2}$$

where g is smooth on R^m, or handling a constraint $f(x) \leq 0$ for such a function. A particularly important case is $f(x) = \lambda_1(x)$, where $g(u_1, \ldots, u_m) = u_1$.

Good insights into this situation are provided through the fact that the functions

$$\Lambda_k(x) = \lambda_1(x) + \cdots + \lambda_k(x) \ \text{ for } \ k = 1, \ldots, m$$

have the envelope representation

$$\Lambda_k(x) = \max_{P \in \mathcal{P}_k} \operatorname{tr}\big(PA(x)P\big), \tag{3}$$

where $\mathcal{P}_k$ is the set of all symmetric $m \times m$ matrices P with rank k such that $P^2 = P$ (i.e., all matrices corresponding to projection mappings onto linear subspaces of R^m of dimension k), and "tr" denotes the trace of a matrix (the sum of its diagonal entries). This fits the pattern of (1) with $\mathcal{P}_k$ as the space S, the "indices" s being matrices P, and $\phi_P(x) = \operatorname{tr}(PA(x)P)$. Obviously one has

$$\lambda_1(x) = \Lambda_1(x), \qquad \lambda_k(x) = \Lambda_k(x) - \Lambda_{k-1}(x) \ \text{ for } \ k = 2, \ldots, m,$$

so f can just as easily be expressed in the form $h\big(\Lambda_1(x), \ldots, \Lambda_m(x)\big)$ for $h(v_1, \ldots, v_m) = g(v_1, v_2 - v_1, \ldots, v_m - v_{m-1})$.

Especially to be noted is the case where the entries $a_{ij}(x)$ of $A(x)$ depend affinely on x, since then $\operatorname{tr}(PA(x)P)$ is affine in x, and it follows that $\Lambda_k(x)$ is convex in x. This implies $\lambda_1(x)$ is convex in x, while $\lambda_2(x), \ldots, \lambda_{m-1}(x)$ are difference-convex (the difference of two convex functions); $\lambda_m(x)$ is actually affine.

In envelope representations of type (3) the index set is a certain compact continuum within a finite-dimensional vector space. Simple discretization would be ill advised, since it would effectively remove the problem from the realm of eigenvalues, where the algebraic foundations are very rich.

Eigenvalue problems also arise for nonsymmetric matrices $A(x)$ and in this case are tougher, because envelope representations aren't at hand. A deeper foray into nonsmooth analysis is required then in identifying the right properties to work with.

For a start on understanding recent work in this branch of nonsmooth optimization, papers of Overton [5] and Overton/Womersely [6] are helpful.

Lagrangian relaxation and decomposition. A major area leading to nonsmooth optimization is that of decomposition schemes for problems of convex type through Lagrange multipliers. These are closely related to Lagrangian relaxation schemes for getting lower bounds to the minimum in problems of nonconvex or combinatorial type.

Starting from a primal problem in which $f_0(x)$ is to be minimized over a subset $X \subset R^n$ subject to constraints $f_i(x) \leq 0$ for $i = 1, \ldots, m$, we suppose that X is compact and that the functions f_i are all smooth. (We stick to inequality constraints for simplicity, and suppose that a feasible solution exists.) The ordinary Lagrangian function associated with this problem is

$$L(x, y) = f_0(x) + y_1 f_1(x) + \cdots + y_m f_m(x) \;\text{ for }\; y = (y_1, \ldots, y_m) \in R^m_+,$$

and the ordinary Lagrangian *dual problem* takes the form

$$\text{maximize } \; g(y) \;\text{ over }\; y \in R^m_+, \;\text{ where }\; g(y) = \min_{x \in X} L(x, y).$$

In general, the optimal value in the primal problem (which is finite under the given assumptions) is bounded below by $g(y)$ for any $y \in R^m_+$; the supremum over all such lower bounds is the optimal value in the dual problem. In some circumstances, notably the case where X and the functions $f_0, f_1, \ldots, f_m$ are all convex, the primal optimal value is known to equal the dual optimal value—the greatest lower bound is then exact. When that holds, and $\bar{y}$ is an optimal solution to the dual, the solutions to the primal are precisely the vectors $\bar{x}$ among those that minimize $L(x, \bar{y})$ over $x \in X$ that happen to satisfy the other primal constraints as well.

Whether the primal problem exhibits convexity or not, there's incentive for possibly trying to solve the dual problem as a means of approaching the primal problem, or at least gaining information about it. This is especially true in situations where for some reason the primal problem is difficult to tackle directly because of the constraints $f_i(x) \leq 0$.

Subproblems in which $L(x, y)$ is minimized over $x \in X$ for some choice of y are called *relaxed* problems because, in comparison with the primal problem, they don't deal with the constraints $f_i(x) \leq 0$ but instead try to reflect them in a modified objective function. The optimal value in such a subproblem is, of course, the dual objective value $g(y)$. Solving a relaxed problem thus produces a lower bound to the primal objective value, which might be very useful. This is important for instance in combinatorial optimization problems where X is a discrete set.

To go from a lower bound $g(y)$ to a better lower bound $g(y')$, one obviously has to employ techniques for making an "ascent" on g. The important thing here is that g may be well be *nonsmooth*. On the other hand, the definition of g furnishes an envelope representation (of minimum instead of maximum type) in which the "indices" are the vectors $x \in X$ and the functions $\phi_x(y) = L(x,y)$ are affine in y. Thus, g is always *concave* in this situation, and the strategies that can be utilized toward maximizing g over R^m_+ are those of convex programming as adapted to handling functions with an envelope representation.

Decomposition methodology puts an additional twist on this. The best-known case of decomposition uses Lagrange multipliers to take advantage of separability. Suppose in the primal problem that the vector $x \in R^n$ has a natural partition into a number of vector or scalar components: let's write $x = (x_1, \ldots, x_r)$, where the components x_k belong to spaces R^{n_k} (with $n_1 + \cdots + n_r = n$). Suppose further that

$$\begin{aligned} f_i(x) = f_{i1}(x_1) + \cdots + f_{ir}(x_r) \ \text{ for } \ i = 0, 1, \ldots, m, \\ X = X_1 \times \cdots \times X_r \ \text{ with } \ X_k \subset R^{n_k}. \end{aligned}$$

The sets X_k could have constraint representations as well, but for now that kind of detail is unnecessary. The Lagrangian then enjoys the special structure

$$\begin{aligned} L(x,y) =& L_1(x_1, y) + \cdots + L_r(x_r, y) \\ & \text{with } \ L_k(x_k, y) = f_{0k}(x_k) + y_1 f_{1k}(x_k) + \cdots + y_m f_{mk}(x_k), \end{aligned}$$

and in the dual problem one has

$$g(y) = g_1(y) + \cdots + g_r(y) \ \text{ with } \ g_k(y) = \min_{x_k \in X_k} L_k(x_k, y).$$

Solving the dual problem amounts therefore to maximizing $g_1(y) + \cdots + g_r(y)$ over $y \in R^m_+$ in a context where every function g_k has its own envelope representation with parameter index $x_k \in X_k$.

Penalty expressions and composite optimization. Penalty terms have most often been viewed as a technical device for dealing with constraints in certain situations, such as within a numerical method. But in applications where caution must carefully be exercised when admitting hard constraints, such as stochastic programming, they have modeling advantages as well, cf. [7], [8].

In proceeding from a problem of minimizing $f_0(x)$ over all $x \in X \subset R^n$ such that $f_i(x) \leq 0$ for $i = 1, \ldots, m$, one can contemplate solving instead a problem of the form

$$\text{minimize } \ f(x) = f_0(x) + \rho_1\big(f_1(x)\big) + \cdots + \rho_m\big(f_m(x)\big) \ \text{ over all } \ x \in X, \tag{4}$$

where each ρ_i is a convex function on R with $\rho_i(0) = 0$. It's helpful in this to allow ρ_i to take on the value ∞, with the understanding that (4) carries the implicit side condition

that $f_i(x)$ should belong to the interval in R where $\rho_i < \infty$. The original problem can be identified with having $\rho_i(u_i) = 0$ when $u_i \le 0$ but $\rho_i(u_i) = \infty$ when $u_i > 0$. The extreme discontinuity of ρ_i in this case underscores the fragility of modeling with hard constraints unless this is strictly necessary.

As alternatives to hard constraints there are rich possibilities. The first that come to mind are classical linear or quadratic penalty terms like $r_i \max\{0, f_i(x)\}$ or $r_i f_i(x)^2$ with $r_i > 0$ as penalty parameter. But ordinary Lagrangian terms $y_i f_i(x)$ fit the picture too, as do augmented Lagrangian terms, which combine multiplier expressions with ultimately quadratic expressions in a piecewise linear-quadratic function ρ_i with y_i and r_i both as parameters. Still other possibilities for ρ_i are barrier expressions or piecewise linear expressions in f_i like $\rho_i\big(f_i(x)\big) = y_i^+ f_i(x)$ when $f_i(x) \ge 0$, $\rho_i\big(f_i(x)\big) = y_i^- f_i(x)$ when $f_i(x) \le 0$, in which the parameter values y_i^+ and y_i^- (with $y_i^+ \ge y_i^-$) specify upper and lower bounds to the range of "shadow prices" to be allowed. Again, such a form of expression might be amalgamated with others.

In general, one can think of the usefulness of convex functions ρ_i that are finite on a certain interval, which is partitioned perhaps into subintervals on which ρ_i has different formulas. Although ρ_i is continuous over the entire interval, its first or second derivatives may not be. Then ρ_i exhibits nonsmoothness, and so too does the function f in (4) that needs to be minimized over X. (Constraints not softened by ρ expressions can be imagined here as incorporated into the specification of X.)

Beyond problems of type (4) there are formats involving composition in a broader manner:

$$\text{minimize} \;\; f(x) = f_0(x) + \rho\big(f_1(x), \ldots, f_m(x)\big) \;\; \text{over} \;\; x \in X, \tag{5}$$

where ρ is a convex but generally nonsmooth function on R^m. All such problem models belong to *composite* optimization, an important branch of nonsmooth optimization.

A nonsmooth function f of the kind in (4), in which the f_i's themselves are smooth and ρ_i's are relatively simple, has many nice properties which can easily be derived and put to use in minimization, for instance in mimicking something like steepest descent. But that's not the only way to go. Techniques of composite optimization focus instead on generating approximations to f in (4) or (5) by preserving ρ while making approximations to each f_i.

Subgradients and subderivatives. Numerical techniques in nonsmooth optimization can be divided roughly into two categories. In direct methods, local properties of the function to be minimized or maximized are developed through variational analysis, including convex analysis, and are utilized in a scheme that emulates well known primal approaches to optimization such as steepest descent, conjugate gradients, and so forth. In indirect methods, modes of approximation arising from Lagrangian functions of one sort or another dominate the scene.

Direct methods depend on the fact that most nonsmooth functions in applications aren't just "bad" functions but have firm handles like envelope representations. We'll

sketch briefly what such a representation in the notation (1) provides. We assume in this that S is a compact space and $\phi_s(x)$ has first-order derivatives in x with these derivatives depending continuously on x and s together. These conditions are trivially satisfied when S is a finite set (in the "discrete topology") and $\phi_s(x)$ is continuously differentiable in s.

First of all, the assumptions guarantee the existence of one-sided directional derivatives that are especially well behaved. At each point $\bar{x}$ and for each vector $\bar{w}$ the limit

$$df(\bar{x})(w) = \lim_{t \searrow 0,\ w \to \bar{w}} \frac{f(\bar{x} + tw) - f(\bar{x})}{t} \tag{6}$$

exists finitely and depends upper semicontinuously on $(\bar{x}, \bar{w})$, in fact continuously on $\bar{w}$. Moreover $df(\bar{x})$, as a function on R^n—called the *subderivative* function for f at $\bar{x}$)—is sublinear, hence convex:

$$df(\bar{x})(w_1 + w_2) \le df(\bar{x})(w_1) + df(\bar{x})(w_2), \qquad df(\bar{x})(\lambda w) = \lambda df(\bar{x})(w) \ \text{ when } \ \lambda > 0.$$

The envelope representation furnishes moreover the formula

$$df(\bar{x})(\bar{w}) = \max_{s \in S_{\bar{x}}} \nabla\phi_s(\bar{x}) \cdot \bar{w} \ \text{ where } \ S_{\bar{x}} = \operatorname{argmax}_{s \in S} \phi_s(\bar{x}). \tag{7}$$

(In other words, $S_{\bar{x}}$ is the set of $s \in S$ at which the maximum in the envelope formula (1) for $f(\bar{x})$ is attained.) Secondly, the closed convex set

$$\partial f(\bar{x}) = \big\{ v \in R^n \big| v \cdot w \le df(\bar{x})(w) \ \text{ for all } \ w \in R^n \big\}, \tag{8}$$

which is called the *subgradient* set for f at $\bar{x}$, is nonempty and compact, and it and depends upper semicontinuously on $\bar{x}$, in the sense that the graph of the set-valued mapping $x \mapsto \partial f(x)$ is closed in $R^n \times R^n$. Furthermore, from the envelope representation one has (with "con" standing for convex hull)

$$\partial f(\bar{x}) = \ \operatorname{con}\big\{ \nabla\phi_s(\bar{x}) \big| x \in S_{\bar{x}} \big\}. \tag{9}$$

From these formulas it's evident that to calculate a subgradient of f at $\bar{x}$, all one has to do is determine a single element $\bar{s} \in S_{\bar{x}}$; then $v = \nabla\phi_{\bar{s}}(\bar{x})$. This requires carrying out the maximization of $\phi_x(\bar{x})$ with respect to $s \in S$, a process which yields the function value $f(\bar{x})$ simultaneously. This which may be easy or hard, depending on the circumstances. In the case of decomposition with Lagrange multipliers, for instance, where y is the variable and x is the "index" and max is replaced by min, it corresponds to solving a family of separate problems in which $L_k(x, \bar{y})$ is minimized with respect to $x_k \in X_k$ for $k = 1, \ldots, r$.

To calculate directional derivatives of f at $\bar{x}$ is harder. If (7) is to be utilized, *all* the elements $s \in S_{\bar{x}}$ may be needed in principle, not just one of them. It's no wonder, then, that direct methods of minimizing a nonsmooth function in terms of an envelope representation

have concentrated on strategies that only require calculating a single subgradient at a time, regarding this as an "expensive" operation, although hardly more expensive than function evaluation. This is the pattern followed in [1]–[4]. Of course, in situations where formulas other than (7) are available for directional derivatives, such as many problem models in composite optimization, where function evaluation may be relatively easy as well, the picture is different and another range of techniques can be brought into play.

Background on the mathematics of subderivatives and subgradients can be found in [7] and the books of Clarke [10] and Rockafellar/Wets [11].

Approximations through generalized Lagrangians. In contrast to direct methods in which a function f is minimized through its subderivatives or subgradients, it's possible often to follow a different path leading to the replacement of the given problem by a sequence of easier problems generated through Lagrangian expressions. The chief domain for this kind of approach is composite optimization, in particular the treatment of penalty expressions.

Consider again a problem expressed in the form (4), where the modeling functions ρ_i on R with values in $(-\infty, \infty]$ are convex, and ρ_i is continuous relative to the closure of the (nonempty) interval D_i where $\rho_i < \infty$. An interesting fact of convex analysis is that for such a function ρ_i there's a dual object, a uniquely determined function ψ_i on R having these same properties, and such that

$$\rho_i(u_i) = \sup_{y_i}\big\{y_i u_i - k_i(y_i)\big\}, \qquad k_i(y_i) = \sup_{u_i}\big\{y_i u_i - \rho_i(u_i)\big\}. \tag{10}$$

In terms of Y_i being the interval of R where $k_i < \infty$, the *generalized Lagrangian function* associated with problem (4) is

$$\begin{aligned}\mathcal{L}(x,y) =& f_0(x) + y_1 f_1(x) + \cdots + y_m f_m(x) - k_1(y_1) - \cdots - k_m(y_m)\\ &\text{for } \ (x,y) \in X \times Y, \ \text{ where } \ Y = (Y_1 \times \cdots \times Y_m).\end{aligned} \tag{11}$$

This isn't some abstraction; the specific form for k_i is well known for the common forms of ρ_i, and in the main cases k_i is smooth on Y_i, in fact typically just quadratic (with $k_i \equiv 0$ as a common special case, the specification of the interval Y_i then being primary). Extension to composite problems in the broader format (5) is easy, but we won't go into that here. An introduction to generalized Lagrangian functions is provided in Rockafellar [9].

The generalized Lagrangian in (11) has, through the first expression in (10), the property that

$$f(x) = \sup_{y \in Y} \mathcal{L}(x,y) \ \text{ for each } \ x.$$

This could be viewed as furnishing another kind of envelope representation for f to which optimization techniques already mentioned could be applied, and indeed it does if Y is compact. A valuable insight, however, is that *the generalized Lagrangian $\mathcal{L}$ well captures all the smoothness that might be used in working with f.* Although f may be a very

complicated function, with its domain divided into numerous regions associated different formulas for $f(x)$, the function $\mathcal{L}$ is simple.

To understand what can be made of this, consider more closely the case where the functions f_i are twice continuously differentiable, the functions k_i are at most quadratic, and the set X is a box (perhaps all of R^n); this already covers a vast array of applications. Then $\mathcal{L}(x, y)$ is twice continuously differentiable in x and y, in particular concave quadratic or affine in y, and the set $X \times Y$ is a box in $R^n \times R^m$. First and second-order conditions for the optimality of $\bar{x}$ in the problem of minimizing $f(x)$ over $x \in X$ can be expressed in terms of first and second derivatives of $\mathcal{L}$ at $(\bar{x}, \bar{y})$, where $\bar{y}$ is a generalized Lagrange multiplier vector.

Analogs of sequential quadratic programming, for instance, can then be envisioned in which, in rawest form, the idea is to generate a sequence of primal-dual pairs (x^ν, y^ν) for $\nu = 0, 1, \ldots$ by taking $\mathcal{L}^\nu(x, y)$ to be the second-order expansion of $\mathcal{L}$ at (x^ν, y^ν), defining f^ν to be the approximation to f corresponding to this expansion, namely

$$f^\nu(x) = \sup_{y \in Y} \mathcal{L}^\nu(x, y) \;\text{ for each }\; x,$$

and then obtaining $(x^{\nu+1}, y^{\nu+1})$ as satisfying the optimality conditions for the subproblem minimizing $f^\nu(x)$ over X. (It would also be possible here to pursue notions of "trust region" in replacing $X \times Y$ iteratively by smaller boxes $X^\nu \times Y^\nu$.) In this kind of scheme $\mathcal{L}^\nu$ is linear-quadratic, and the problem of $f^\nu(x)$ over X is said to be one of *extended linear-quadratic programming*. Actually, one gets (for the same functions ρ_i) that

$$f^\nu(x) = f_0^\nu(x) + \rho_1\big(f_1^\nu(x)\big) + \cdots + \rho_m\big(f_m^\nu(x)\big)$$

where f_0^ν is the second-order expansion of $\mathcal{L}(x, y^\nu)$ in x at x^ν and, for $i = 1, \ldots, m$, f_i^ν is the first-order expansion of f_i at x^ν.

Although problems of extended linear-quadratic programming may still display rampant nonsmoothness in the primal objective function f^ν, they have their own characteristics which facilitate computation in other ways. When convexity is present, for example, they can be approached in terms of calculating a saddle point of $\mathcal{L}(x, y)$ with respect to $(x, y) \in X \times Y$. This is a subject in which many new computational ideas have recently been developed. See for instance [12] and its references.

Parametric optimization.

Yet another important source of nonsmoothness in optimization is found in decomposition schemes where a problem's variables are divided into "easy" and "hard." Suppose that the ultimate goal is to

$$\begin{aligned} &\text{minimize } \; f_0(w, x) \;\text{ over all }\; (w, x) \in W \times X \\ &\text{with } \; f_i(w, x) \leq 0 \;\text{ for }\; i = 1, \ldots, m, \end{aligned} \tag{12}$$

where $W \subset R^d$ and $X \subset R^n$. (Broader problem models on the order of (4) or (5) could be regarded in the same light.) Imagine that w stands for the "easy" variables, in the sense that for any fixed $x \in X$ it's relatively easy to compute

$$\begin{aligned} f(x) = \ &\text{minimum of } \ f_0(w,x) \ \text{ in } \ w \in W \\ &\text{subject to } \ f_i(w,x) \le 0, \quad i = 1, \ldots, m. \end{aligned} \tag{13}$$

Then there is the residual problem of minimizing $f(x)$ over $x \in X$. If an optimal solution $\bar{x}$ can somehow be found for that, an optimal solution to the underlying problem (12) will be given by $(\bar{w}, \bar{x})$ for any $\bar{w}$ solving (13) for $x = \bar{x}$.

The main obstacle in this situation is, of course, the nonsmoothness of f. The hope is that information pertinent to minimizing f over X can be gleaned from calculations in the subproblems (13) for various choices of x. When W, X, and all the functions f_i are convex, f is at least convex, and special techniques can be used. It may be possible to proceed by dualizing (13) to obtain an envelope representation for f, which for instance is the approach of Benders decomposition. In general, though, an envelope representation may not be obtainable. This kind of nonsmoothness is then the most difficult to handle, because f doesn't have nice subderivatives and subgradients as described above in terms of such a representation. In certain cases such as those reviewed by Gauvin [13], Lagrange multipliers provide relatively accessible knowledge about directional derivatives. More generally the concepts of subderivative and subgradient have robust extensions to such a context (see [10] and [11]), but their utilization in effective methods of computation has hardly yet been explored.

Nonsmoothness of other orders. The discussion has revolved around nonsmoothness in a function f that one wishes to minimize, but other forms of nonsmoothness arise in areas of optimization where optimality conditions from one problem are introduced as constraints in another problem, or simply when attempts are made to solve first-order optimality conditions as if they resembled a system of nonlinear equations. This is the subject of *generalized equations*.

As a key example, a problem of the broad type (4), which covers traditional optimization problems as the case where ρ_i is 0 on $(-\infty, 0]$ or $[0, 0]$ but ∞ elsewhere, has first-order optimality expressible in terms of the generalized Lagrangian (11) by

$$-\nabla_x \mathcal{L}(\bar{x}, \bar{y}) \in N_X(\bar{x}), \qquad \nabla_x \mathcal{L}(\bar{x}, \bar{y}) \in N_Y(\bar{y}), \tag{14}$$

where $N_X(\bar{x})$ is the normal cone to X at $\bar{x}$ and $N_Y(\bar{y})$ is the normal cone to Y at $\bar{y}$. When X and Y are boxes, for instance, these normal cone conditions reduce to sign conditions on the components of $\bar{x}$ and $\bar{y}$ and the partial derivatives of $\mathcal{L}$. The pairs $(\bar{x}, \bar{y})$ are the generalized Kuhn-Tucker points associated with the problem.

Consider in this vein the sets

$$\begin{aligned} G = \big\{ &(x,y,u,v) \in R^n \times R^n \times R^n \times R^n \big| \\ &\qquad -\nabla_x \mathcal{L}(\bar{x}, \bar{y}) + v \in N_X(\bar{x}), \ \nabla_x \mathcal{L}(\bar{x}, \bar{y}) - u \in N_Y(\bar{y}) \big\}, \\ M = \big\{ &(x,y,u,v) \in R^n \times R^n \times R^n \times R^n \big| u = 0, \ v = 0 \big\}. \end{aligned}$$

Trying to determine $(\bar{x}, \bar{y})$ can be viewed as trying to find an element $(\bar{x}, \bar{y}, \bar{u}, \bar{v}) \in G \cap M$. The idea comes up then of devising algorithms patterned after ones that might work if G were a smooth manifold given by nice, nondegenerate equations. For instance, one can imagine creating a sequence of local first-order approximations G^ν to G at points $(x^\nu, y^\nu, u^\nu, v^\nu)$, where in basic concept $(x^{\nu+1}, y^{\nu+1}, u^{\nu+1}, v^{\nu+1})$ is determined as a point of $G^\nu \cap M$ in a Newton-like scheme.

The challenge here is that G isn't just a smooth manifold, and doesn't have first-order approximations in the sense of classical linearizations. It's a *nonsmooth* manifold, moreover of a kind requiring an advanced form of nonsmooth analysis. But actually the properties of G are convenient and attractive nevertheless. Natural and simple first-order approximations do exist. In particular, these can be obtained through linearizing $\nabla\mathcal{L}$, i.e., working in effect with quadratic approximations to $\mathcal{L}$ as already discussed.

An introduction to the methodology being developed for solving nonsmooth equations is furnished in Pang/Qi [14].

References

1. K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Math. 1133, Springer-Verlag, Berlin, 1985.
2. N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985.
3. J-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms, I & II*, Springer-Verlag, Berlin, 1993.
4. J. Zowe, "The BT-algorithm for minimizing a nonsmooth functional subject to linear constraints," in *Nonsmooth Optimization and Related Topics*, F.H. Clarke, V.F. Demyanov and F. Giannessi (eds.), Plenum Press, 1989, 459–480.
5. M. L. Overton, "Large-scale optimization of eigenvalues," SIAM J. Optimization 2 (1992), 88–120.
6. M. Overton and R. S. Womersley, "Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices," Math. Programming (1993).
7. R. T. Rockafellar and R. J-B Wets, "A Lagrangian finite generation technique for solving linear-quadrataic problems in stochaseic programing," Math. Programming Studies 28 (1986), 63–93.
8. R. T. Rockafellar and R. J-B Wets, "Linear-quadratic problems with stochastic penalties: the finite generation algorithm," in *Stochastic Optimization*, V. I. Arkin, A. Shiraev and R. J-B Wets (eds.), Lecture Notes in Economics and Math. Systems 255, Springer-Verlag, Berlin, 1987, 545–560.
9. R. T. Rockafellar, "Lagrange multipliers and optimality," SIAM Review 35 (1993), 183–238.

10. F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983; republished by the Centre de Recherches Mathématiques, Université de Montréal, C.P. 6128A, Montréal, Quebec, Canada.
11. R. T. Rockafellar and R. J-B Wets, *Variational Analysis*, forthcoming book.
12. C. Zhu and R. T. Rockafellar, "Primal-dual projected gradient algorithms for extended linear-quadratic programming," SIAM J. Optimization 3 (1993), 751–783.
13. J. Gauvin, "Directional derivative for the value function in mathematical programming," in *Nonsmooth Optimization and Related Topics*, F. H. Clarke, V. F. Demyanov and F. Giannessi (eds.), Plenum Press, 1989, 167–183.
14. J-S. Pang and L. Qi, "Nonsmooth equations: motivation and algorithms," SIAM J. Optimization 3 (1993), 443–465.