SUBGRADIENTS AND VARIATIONAL ANALYSIS

R. T. Rockafellar

57th Meeting of the Indian Mathematical Society*

Aligarh, December 27–30, 1991

Abstract. The study of problems of maximization or minimization subject to constraints has been a fertile field for the development of mathematical analysis from classical times. In recent decades, convexity has come forward as an important tool, and the geometry of convexity has been translated into notions of directional derivatives and subgradients of functions that may not be differentiable in the ordinary sense. Now there has emerged a form of analysis able to deal robustly, even in the absence of convexity, with the phenomena of nonsmoothness that arise in variational problems.

The Need for Variational Analysis

Differential calculus has been so successful in treating a variety of physical phenomena that mathematics has long relied on differentiable functions as the main tools of analysis. The domain of such a function, in a finite-dimensional setting, is typically an open subset of $I\!R^n$ or of a differentiable manifold such as might be defined by a system of equations in $I\!R^n$ and coordinatized locally by $I\!R^d$ for some $d < n$. Many of the systems and phenomena that have come under mathematical scrutiny in recent decades, however, are of interest especially at their frontiers of feasibility. They involve functions and mappings whose domains may be closed sets with very complicated boundaries, expressed often by numerous inequalities as well as equations. Behavior around boundary points of these domains is seen as crucial, but it cannot well be investigated without a development of ideas beyond the customary framework.

A major source of this trend lies in the fact that mathematical models are being used more and more for *prescriptive* as well as merely *descriptive* purposes. Nowadays one seeks not only to describe what happens in the world but to influence or improve the way it happens. New subjects have been created like optimization theory, control theory, and viability theory, which are heavily involved with finding the extremes of what may be possible under given circumstances. This has stemmed from an increasing preoccupation

* The written text of this lecture has been adapted from the introduction of the forthcoming book *Variational Analysis* by R. T. Rockafellar and R. J-B Wets.

with problems in engineering design, operations management, economics, statistics, and biology, alongside of those in the physical sciences.

Computers have made it possible to cope with a vast array of such problems, too complex to have been taken on in the past. They have opened up new applications of mathematics and at the same time have caused a shift in attitude, where "closed form" solutions are removed from the pedestal as the compelling ideal, and structure conducive to obtaining numerical answers is emphasized instead. Whether a solution is expressed by a formula in classical terms or by an algorithm for computation makes little difference *if the tools are available for understanding the nature of the solution and how it depends on the data*.

Variational analysis, which has grown from these challenges, extends classical analysis by admitting a much wider class of functions and incorporating them into a "subdifferential" calculus that continues to utilize standard properties when possible but is not limited by them. Smoothness, a term referring to continuous differentiability and its graphical counterparts, no longer plays a dominating role. Instead the spotlight is on one-sided tangents, one-sided derivatives, and "subgradients." This part of the subject is often called *nonsmooth* analysis, but variational analysis covers other ground too and is not concerned just with nonsmoothness.

Set-valued mappings and extended-real-valued functions emerge as central objects of study for generalizations of continuity, convergence and approximation as well as differentiability. The measurability theory of set-valued mappings and the special measurability properties of extended-real-valued "integrands," which are essential to the formulation of many problems involving integrals over time or expectations in probability, are an important part of variational analysis also. Operations of maximization and minimization, because of their significance in identifying extremal behaviors and describing properties of boundaries, are elevated to a status comparable to that of integration in being used systematically to define functions and mappings and to express their derivatives, rather than merely to state problems. Characteristically, there is a strong tie to geometry throughout the subject, but the modes of geometric thinking are often different from the long-familiar ones.

A core idea is that of analyzing local variations or perturbations of a mathematical object, like a set or mapping, when these variations or perturbations are liable to run up against crucial restrictions, or are admitted only relative to side conditions which may be far from transparent in their effects and require close study in themselves. A basic purpose lies in describing the circumstances in which a given function achieves an extreme value over a given set, and in answering questions of the stability and sensitivity to change exhibited by the extreme value or the points where it is attained. A further purpose is to set the mathematical stage for techniques of computing such points.

Constraints and Nonsmoothness

In a context of searching for extremes, the concept of *constraints* is fundamental. Constraints are the conditions that define the set of elements over which a particular search takes place. They may have many forms, but quite commonly they involve restrictions on the values of a collection of functions of variables x_j, which are interpreted as the coordinates of a vector $x = (x_1, \ldots, x_n)$.

A typical example is a set $C \subset I\!R^n$ is defined by

$$C = \left\{ x \in X \,\middle|\, f_i(x) \le 0 \text{ for } i \in I_1 \text{ and } f_i(x) = 0 \text{ for } i \in I_2 \right\} \tag{1}$$

where X denotes another subset of $I\!R^n$, perhaps $I\!R^n$ itself, and I_1 and I_2 are general index sets for a collection of real-valued functions. The conditions $f_i(x) = 0$ are then called *equality constraints*, while the conditions $f_i(x) \le 0$ are *inequality constraints.* Each f_i is a *constraint function.* (The condition $x \in X$ in this example is referred to as a constraint too, even though it is on an abstract level until more is specified about X.)

In classical mathematical models in this format, there are typically only a few, simple constraints: I_1 and I_2 are small, finite sets. In fact there are usually only equality constraints, and X is open. The constraint functions are differentiable, and their gradients satisfy conditions of linear independence. This gives C the character of a *differentiable manifold.* Even when inequality constraints do come into play, they are relatively elementary. For instance, one might think of a cube in $I\!R^3$ or the intersection of two balls. Then, although C may not itself be treatable as a differentiable manifold, it can be decomposed into a manageable number of readily identifiable pieces, each of which is a differentiable manifold in the form of a curve, surface, or open region. These pieces can be analyzed separately.

In contrast, the problems seen in applications today may involve sets C described by conditions (1) in which the inequality constraints are dominant and could number in the thousands, if not millions—often as a result of discretization in time, space, or probability. There can be far more of such constraints than variables x_j, although the number of those can be enormous too. Approaches based only on "smoothness" are then totally inadequate.

At any particular point x of C, some of the inequality constraints in (1) can be *active* (satisfied as equations $f_i(x) = 0$) while others are *inactive* (satisfied as strict inequalities

$f_i(x) < 0$). Quite apart from the large numbers involved, there is usually no easy way to determine which combinations of active versus inactive constraints actually do occur. Even if there were, one would not necessarily get a nice decomposition of C into differentiable manifolds, due to possible breakdowns of linear independence among constraint gradients. Anyway, there would be too many subsets in the decomposition to make a case-by-case analysis reasonable.

To look at this from a different angle, suppose that C is specified by inequality constraints alone:

$$C = \{ x \in I\!R^n \,|\, f_i(x) \le 0 \text{ for } i \in I \} \tag{2}$$

where the f_i's are smooth and the index set I is finite but large. Suppose in addition that the interior of C consists of the points x satisfying $f_i(x) < 0$ for all $i \in I$, and that C is the closure of its interior. The boundary of C might then be thought of as a kind of "nonsmooth surface." In general there would be no way to study it in separate pieces, and one would hope instead to find some direct approach.

Note by the way the inherent asymmetry in the boundary set in this example (as imagined in more than two dimensions). It has all its "creases" and "corners" on the side corresponding to the exterior of C, none on the side corresponding to the interior. For this reason it is better studied jointly with C than as a hypersurface all on its own. Such one-sided treatments of mathematical structure are common in variational analysis.

The pictures created by inequality constraints are seen also when functions are defined in terms of the operations "max" and "min." Suppose we have

$$f(x) = \max_{i \in I} f_i(x), \tag{3}$$

again with f_i smooth. This formula means that the value of f at a point x is taken to be the highest of all the values $f_i(x)$ as i ranges over I. Regardless of the degree of differentiability of the f_i's, the function f is unlikely to be smooth. There is a strong resemblance between this situation and that of a "nonsmooth set," and this is hardly an accident, because the graph of f is the boundary of a set D defined much as C was in (2):

$$D = \{ (x, \alpha) \in I\!R^n \times I\!R \,|\, f_i(x) - \alpha \le 0 \text{ for } i \in I \}.$$

This example brings out a very important connection between nonsmooth sets and nonsmooth functions, as induced by inequality constraints in one direction and by "max" or "min" in the other. It is often worthwhile in variational analysis to pass between the two points of view. In the study of f defined by (3) one can apply geometric concepts most profitably to the set D, the so-called *epigraph* consisting of the points in $I\!R^{n+1}$ lying on or above the graph of f, rather than to the graph itself. In the other direction, the set C in (2) can be expressed in terms of f as

$$C = \{x \in I\!R^n \mid f(x) \leq 0\}. \tag{4}$$

The trick of aggregating the constraints defining C into a single "nonsmooth constraint" by means of (3) is attractive for a number of purposes. But of course it cannot be used to much effect unless techniques are available for working with the functions f that are thereby created.

The virtues of studying sets by way of inequality constraints, and functions by way of their epigraphs, were first appreciated in the special context of *convex* analysis, for which [1] serves as a reference. A subset C of $I\!R^n$ is said to be *convex* if it includes for every pair of points the line segment that joins them, or in other words, if for every choice of $x_0 \in C$ and $x_1 \in C$ the point $x_0 + \tau(x_1 - x_0) = (1-\tau)x_0 + \tau x_1$ belongs to C also for all $\tau \in (0,1)$. A function f on a convex set C is said to be *convex* relative to C if for all x_0 and x_1 in C it satisfies

$$f\big((1-\tau)x_0 + \tau x_1\big) \leq (1-\tau)f(x_0) + \tau f(x_1) \text{ for all } \tau \in (0,1).$$

The closed convex sets in $I\!R^n$ turn out to be the sets representable by *linear* inequalities as in (2), but with a possibly infinite index set. The convex functions on $I\!R^n$ can be studied geometrically through the fact that their epigraphs are convex sets. When the epigraph of such a function f is closed, a property termed *lower semicontinuity*, its representation by a system of linear inequalities corresponds to a representation of f as in (3) by a possibly infinite collection of f_i's, each of which is a linear function plus a constant.

The theory of convexity is thus inevitably concerned with nonsmoothness. The potential lack of differentiability notwithstanding, convexity is a natural assumption in economics, for instance, where smoothness assumptions are hard to justify axiomatically. Not surprisingly in this connection, mathematical models in economics and related areas are especially rich in inequality constraints. Goods can be present only in nonnegative quantities. Resource stocks may be under-utilized if not needed, but no quantity larger than is available can be drawn into production.

Inequality constraints are frequently seen in engineering in the form of bounds placed on certain variables. They also occur in fields like physics and chemistry, although little attention was paid to this in the past. A chemical system such as a planetary atmosphere, oil reservoir, or human blood, can in some situations be modeled by "states" that specify the current amounts of various chemical species in a number of different phases. A state is then a point in some space $\mathbb{R}^n$. Restrictions on which states are chemically realizable are given by mass balance equations and laws of interaction, but also by the simple fact that no species can occur in an amount less than zero. Zero amounts, on the other hand, are quite possible: a particular substance may, for instance, be present only in one of several possible liquid or solid phases. The set C consisting of all realizable states is determined therefore by numerous inequality constraints (nonnegativity) combined with a number of equations.

To find the state of equilibrium in such a chemical system, one would be obliged to minimize over this complicated set C a certain energy function f. The function f in this case (Gibbs free energy) happens not to be differentiable everywhere on C or for that matter even to have a natural extension beyond C. This poses a serious obstacle in understanding the nature of equilibrium, at least within the confines of classical methodology. The obstacle is all the greater in any attempt to see how such a system might evolve over time in obedience to laws of dynamics. Variational analysis, however, aims at the treatment of just such situations in addition to the classical ones.

Note that the study of f and C in this application could serve not only for characterizing the equilibrium state of the chemical system. It could lead to schemes for calculating it. Here we see a modern twist to a venerable topic, that of expressing the equilibrium states of a system as the ones that minimize an energy function—a so-called *variational principle*. Where previously this was mainly an interesting interpretation that could be given for some forms of equilibrium, it can now serve also not only as a basis for computation but in extending the conditions for equilibrium to systems subject to more complicated constraints than the ones formerly considered.

Variational principles in this general vein are one of the prime motivations for the development of the new forms of analysis. Besides physics and chemistry, they appear for instance in economics. The static equilibrium states of an economy have been characterized under certain assumptions in terms of producers maximizing their profits and consumers maximizing their "utility."

Optimization

Variational principles are often associated with descriptive models of phenomena, but prescriptive models involve the very same theoretical challenges, from a mathematical standpoint. For both kinds of models, the notion of an optimization problem has proven to be valuable. We shall rely on this concept even though the term "optimization," in apparently referring to a search for the "best" among possibilities, is not linguistically ideal for all situations.

An *optimization problem* (in the sense of minimization) is specified by a set C and a real-valued (or extended-real-valued) function f on C, called the *objective function*. The elements $x \in C$ are called the *feasible solutions* to the problem. The ones that minimize f relative to C, if any, are called *(globally) optimal solutions*, and the greatest lower bound (not necessarily finite) for f relative to C is called the *optimal value*. (Minimization could be replaced here by maximization, but of course every problem of maximization can be converted to one of minimization by a change of sign.)

The double use of the word "solution" in this definition may seem odd, but it reflects a fundamental two-stage structure in many mathematical models. The set C is usually defined by constraints within some larger, standard space like $I\!R^n$. Not only may these constraints be complicated, they may in some cases even be inconsistent. This is not merely due to poor formulation. An important question may be whether there is any x at all that satisfies the full collection of constraints, and if so, how one may be determined numerically. (Such a point might in particular be desired to initiate an algorithm for minimizing f over C.) An x belonging to C can therefore appropriately be viewed as a type of solution at a preliminary level. "Feasible," in referring to the fulfillment of constraints, means roughly the same as "admissible" or "realizable," or in certain dynamical settings, "viable."

Whether the ultimate interest in an optimization problem resides in the optimal value (a number), or the optimal solutions (distinguished elements of the feasible set), or both, would depend on circumstances. A particular optimization problem is by definition always focused on just one of the operations of maximization or minimization, however.

Extended-real-valued functions enter the territory of optimization theory by several routes, but one of the most important is the use of "infinite penalty" representations of constraints. To explain this briefly, consider again a problem of minimizing f over C, and define the function $\bar{f} : I\!R^n \to \overline{I\!R} = [-\infty, \infty]$ by

$$\bar{f}(x) = \begin{cases} f(x) & \text{when } x \in C, \\ \infty & \text{when } x \notin C. \end{cases}$$

Then the given problem becomes that of minimizing $\bar{f}(x)$ over all $x \in I\!R^n$, since the points where $\bar{f}$ has the value ∞ are the least interesting and in effect are excluded from consideration. An optimization problem in n variables subject to various constraints can thus be identified with a single function $\bar{f}$ on $\mathbb{R}^n$.

Of course, in passing to an extended-real-valued function in this manner another step is taken away from classical analysis. In particular, the graph of $\bar{f}$ can no longer be the basis of geometric thinking, since it would be a subset of $\mathbb{R}^n \times \overline{\mathbb{R}}$ instead of $\mathbb{R}^n \times \mathbb{R}$, where $\overline{\mathbb{R}} = [-\infty, \infty]$, and $\mathbb{R}^n \times \overline{\mathbb{R}}$ is not a vector space. But the epigraph $\{ (x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \,|\, \alpha \geq \bar{f}(x) \}$ remains as a subset of a vector space and can fill the geometric gap.

Many complications can arise in a careful investigation of optimal solutions, even in a "smooth" case where there are no constraints. There can well be more than one optimal solution, in fact infinitely many of them, not necessarily isolated from each other. Misleading "locally optimal" solutions can be arbitrarily close by. Still, in following a function f downward, as in generating a sequence of points $x_k \in C$ for which the values $f(x_k)$ are decreasing, there is no guarantee that an optimal solution will eventually be approached, even if the sequence of numbers $f(x_k)$ does itself converge to the optimal value.

One of the nicest consequences of convexity, when it is present, is that troubles with local versus global optimality disappear. With a certain "strict" convexity, multiple optimal solutions are eliminated too. As a matter of fact, convexity conditions are virtually the *only* practical way of safeguarding against the difficulties indicated—practical in the sense that a criterion can be checked in terms of the given features of a problem in advance of computations, the results of which might otherwise be hard to interpret rigorously. There is consequently a high premium on being able to recognize the presence of convexity, and a well developed apparatus does exist for this purpose. *While in classical analysis the division between linear and nonlinear is the principal watershed, in variational analysis the division between convex and nonconvex has this prominence.*

Variational Inequalities

The solutions to an optimization problem or variational principle cannot be analyzed and interpreted without first characterizing them in terms of some kind of necessary conditions and sufficient conditions. This is all the more a prerequisite for the development of numerical techniques for computing solutions.

For a problem of minimizing a smooth function f over $I\!R^n$ without constraints, a locally optimal solution $\bar{x}$ must satisfy $\nabla f(\bar{x}) = 0$. The calculation of locally optimal solutions can thus be identified to some extent with the calculation of solutions to a possibly nonlinear equation. As a matter of fact, much of the methodology in numerical analysis for solving linear and nonlinear equations $M(\bar{x}) = 0$ for a vector-valued mapping M is concentrated precisely on cases where $M = \nabla f$ for some function f. Many of the applications are to finite-dimensional discretizations of operator equations in function spaces, where a variational principle may be involved. For instance, a partial differential operator of elliptic type can be expressed as the gradient of a certain kind of convex integral functional, so that partial differential equations with such operators have this quality.

For minimization problems with constraints, the characterization of optimal solutions is much more challenging, especially when large numbers of inequality constraints may be involved. To cope with the difficulties, new levels of abstraction are required so that the essential features of a situation or class of numerical methods can be surveyed apart from a mass of distracting details.

An example is the concept of a "variational inequality." In the case of a nonempty, closed, convex set $C \subset I\!R^n$ and a continuous mapping $M : C \to I\!R^n$, a point $\bar{x}$ is said to satisfy the *variational inequality* for C and M if

$$\bar{x} \in C \text{ and } \langle M(\bar{x}), x - \bar{x} \rangle \geq 0 \text{ for all } x \in C. \tag{5}$$

Here $\langle v, w \rangle$ denotes the standard inner product in $I\!R^n$.

This condition gives some insight into the modern ways of thinking in variational analysis. When the mapping M in a variational inequality is the gradient ∇f of a smooth, real-valued function f, the condition is related to local minimization because the negative of the vector $M(x) = \nabla f(x)$, if nonzero, points in the direction of greatest decrease of f. This provides the simplest case for interpretation, but applications involving "minimax" problems, or involving no minimization or maximization at all, also occur.

If $\bar{x}$ belongs to the interior of C, the variational inequality condition is just the equation $M(\bar{x}) = 0$. If $\bar{x}$ is on the boundary, however, the condition has a one-sided quality: it restricts the angle that $-M(\bar{x})$ is allowed to make with difference vectors $x - \bar{x}$ corresponding to the various points x of C other than $\bar{x}$. The angle cannot be acute.

The formulation of the variational inequality in this manner, without just saying whether $\bar{x}$ is to be on the boundary or in the interior of C, is of the essence. In practice there is usually no way of knowing in advance which case will prevail, and it would be unwieldy to have to deal with the matter in terms of a list of special alternatives, each dependent on the ultimate location of $\bar{x}$.

Variational inequalities were first introduced in an infinite-dimensional setting for the sake of applications to partial differential operators, which may be the gradient mappings associated with certain integral functionals, as already noted. They make it possible there to handle problems with one-sided boundary conditions or obstacles, as for instance a membrane of minimal area stretched around a given object and fastened in certain places. The variational inequality is then a "partial differential inequality" which can be viewed as the optimality condition corresponding to a variational principle involving inequality constraints. A finite-dimensional variational inequality can correspondingly arise through discretization for purposes of computation, just as in the case of a classical partial differential equation being converted into the numerical solution of an equation $M(\bar{x}) = 0$.

Perturbations

Critical to the study of almost all mathematical models—and optimization problems are no exception—is the issue of sensitivity with respect to changes in input data and various parameters. This too is a prime stimulus to developments in variational analysis, because it quickly leads beyond the range of classical theory, a fact not always appreciated enough by those working with applications of minimization. An impression often gained from traditional applied mathematics is that any well formulated and physically well motivated problem one may wish to solve will, under palatable assumptions that could be gone into if necessary, have a unique solution which reacts in only minor ways to minor perturbations. Such may be true in some degree for problems built up with standard operations for which the continuity properties have long been established, but minimization is not an operation in that class, as already emphasized earlier in this discussion.

Approaching the issue systematically, let us consider a minimization problem in variables $x_1, \ldots, x_n$ that depends on other variables $u_1, \ldots, u_d$. Specifically, let us suppose that for each vector u in a certain set $U \subset \mathbb{R}^d$ we want to look at the problem

$$\begin{aligned} &\text{minimize } f_0(u,x) \text{ over all } x \in C(u) \text{ (for fixed } u\text{), where} \\ &C(u) = \big\{ x \in X \,\big|\, f_i(u,x) \le 0 \text{ for } i \in I_1,\ f_i(u,x) = 0 \text{ for } i \in I_2 \big\}. \end{aligned} \tag{6}$$

Let $p(u)$ denote the optimal value in this problem for a given u, and let $P(u)$ denote the corresponding set of optimal solutions. What can be said about the way that $p(u)$, $P(u)$

and $C(u)$ vary with u?

It is unlikely, except in quite special circumstances, that $p(u)$ will be a differentiable function of u, no matter how smooth f_0 and the constraint functions f_i are, even if $X = \mathbb{R}^n$. Moreover, p might sometimes fail to be continuous. Still, there are many situations where an understanding of possible rates of change of p, to the extent that these might be captured by a notion of differentiation more flexible than the standard one, could be very valuable. For example, there are applications where a function defined like p can enter into a different optimization problem as its objective function or as one of the constraint functions. This is especially seen in schemes for decomposing large problems into smaller ones.

The study of the optimal solution set $P(u)$ for the problem in (6) brings up other difficulties which must be dealt with squarely. As noted, there is frequently no way to guarantee that $P(u)$ always consists of a single element without making restrictions that could be severe, unrealistic, or impossible to check. A multiplicity of optimal solutions may well occur for at least some choices of u. On the other hand, $P(u)$ could be empty for other choices of u, not merely because of a failure of existence criteria for optimal solutions, but because constraints could become inconsistent. It is essential, therefore, to treat P as a *set-valued* mapping and to pay attention to the boundary of its "effective" domain: the set of vectors u such that $P(u) \neq \emptyset$. This domain would typically be a closed set with a boundary whose treatment might be problematical. Again there is incentive for developing notions of generalized differentiability, not to mention continuity, that adapt well to the desired applications.

Such issues in the study of the optimal solution set $P(u)$ are central likewise in coming to grips with the feasible solution set $C(u)$ as defined parametrically in (6). Still further in this vein, the variational inequality in (6) could depend on parameters: instead of $M(x)$ for $x \in C$ one could have $M(u, x)$ for $x \in C(u)$. It would then have a solution set $S(u)$ dependent on u, and the effects on $S(u)$ of "errors" in u could be of keen interest.

Analysis of Set-Valued Mappings

To some extent, therefore, variational analysis necessarily takes on the character of what might be called *set-valued* analysis. Set-valued mappings must somehow be treated on a par with ordinary single-valued mappings. But set-valued mappings are "nonsmooth" in the more profound sense that in the past there never was any calculus for them, not the slimmest recipe for derivatives. Only in fairly recent times has a substantial theory even of continuity and semicontinuity of set-valued mappings been put together. That theory must inevitably play a part in any broadening of the calculus to deal with the challenges already mentioned, but it can also serve purposes of its own.

In working with $P(u)$ and $C(u)$, for example, one needs to know what meaning to assign to the convergence of a sequence of sets to a limit set. Concepts of set convergence turn out also to be the key to a robust theory of approximation of one optimization problem by another, such as can be put to good use in the development of numerical methods. They are essential too in understanding the perturbations of a given problem, and—on the level of mathematical modeling in complex situations where asymptotics necessarily take over—sometimes even in constructing the natural "limit problem" that ought to be solved.

Such is the case with the infinite-dimensional technique known as stochastic homogenization, for instance. An optimization model or variational principle may be set up that involves a porous material with random holes of various sizes, but where a specific representation of all the holes by inequality constraints, say, would be out of the question. On the other hand, existing physical theory may not provide a clear substitute for such a representation. In that case one has to try to identify through convergence the correct asymptotic model, involving a kind of averaging, but the limit that must be taken does not fit a traditional mold.

The basics of set convergence lead to the companion notion of "epi-convergence" of functions, which corresponds to set convergence of epigraphs. Such nonclassical function convergence has strong appeal in many applications involving optimization in its broad sense, but also in purely theoretical investigations. The role of set convergence and epi-convergence in generalized differentiation is also crucial.

The need for looking beyond an ordinary framework of point-to-point mappings and continuity is illuminated from a different side by the case of mappings that are monotone in the following sense. A mapping T that associates with each point $x \in I\!R^n$ a subset $T(x) \in I\!R^n$ is called *monotone* if

$$\langle v' - v,\, x' - x \rangle \geq 0 \text{ whenever } v \in T(x) \text{ and } v' \in T(x').$$

It is a *maximal* monotone mapping if there is no monotone mapping $T' \neq T$ with $T'(x) \supset$

$T(x)$ for all x.

The reason for mentioning monotone mappings here is that although they are often merely single-valued in the sense of $T(x)$ consisting of just a single vector v, and quite generally are single-valued "almost everywhere," their full development depends on abandoning single-valuedness. This is because many of the constructions utilized in applications of such mappings, like taking inverses, destroy single-valuedness yet preserve monotonicity. Furthermore, the basic geometry of monotonicity becomes clear only through consideration of the idea that every monotone mapping can be enlarged to a maximal one, which can fail to be single-valued even though single-valuedness may have been present initially.

In the one-dimensional case there is a close resemblance of the graph of T to that of a function from $\mathbb{R}$ to $\mathbb{R}$. To get the graph of a function, it would only be necessary to replace each of the vertical segments by a single one of its elements; then T would still be monotone but no longer maximal monotone. The higher-dimensional analogs, although harder to visualize, are similar, in that a maximal monotone mapping T can be described rather generally as coming from a single-valued one with domain and range in $\mathbb{R}^n$ by filling in, at points of discontinuity, between the different limit values obtainable by approaching from different directions. Multivaluedness is thus introduced at such points. The reward for allowing it is that the resulting graph set turns out to be a connected n-dimensional manifold in $\mathbb{R}^n \times \mathbb{R}^n$.

In higher dimensions monotonicity is a kind of generalization of *positive semidefiniteness*. Indeed, if $T(x)$ consists for each x of a single vector $v = Ax + a$, where A is a matrix in $\mathbb{R}^{n \times n}$ (not necessarily symmetric) and a is a given vector in $\mathbb{R}^n$, one has T monotone if and only if $\langle x, Ax \rangle \geq 0$ for all x, and T strictly monotone if and only if $\langle x, Ax \rangle > 0$ for all $x \neq 0$.

While the study of monotone mappings inevitably requires a push beyond classical analysis, some of the necessary innovations can well be anticipated. Addition of monotone mappings can be developed in a natural way along with other operations. Approximations can be carried out in terms of closeness of the graph sets. A theory of limits of mappings—set-valued as well as single-valued—is thereby suggested which is not describable in the usual terms of pointwise convergence of functions, much less uniform convergence, but through a geometrical notion of *graph* convergence.

Monotone mappings are a key ingredient in numerical procedures for solving variational inequalities and optimization problems in a setting of convexity. Many problems can be represented in terms of finding a point $\bar{x}$ such that $0 \in T(\bar{x})$, where T is maximal monotone.

The development of procedures for solving $0 \in T(\bar{x})$, whether T is maximal monotone or not, is a step beyond the numerical analysis of a nonlinear equation $0 = T(\bar{x})$ for a smooth single-valued mapping T, but for many applications it is quite analogous. There is strong incentive therefore in keeping the parallels between set-valued and ordinary single-valued mappings as close as possible from a theoretical perspective. As for generalized differentiation of T at $\bar{x}$, one aim can then be to characterize convergence properties of algorithms just as may be done in the single-valued, smooth case of T in terms of the Jacobian matrix for T at $\bar{x}$.

Set-valued mappings are of interest in numerical work for more general reasons as well. Quite commonly an algorithm can be described as generating a sequence $\{x_k\}_{\nu\in\mathbb{N}}$ through a procedure of the form $x_k \in S(x_{k-1})$ that starts from a point x_0, the goal being convergence to a point $\bar{x}$ such that $\bar{x} \in S(\bar{x})$. Here the mapping S might be replaced by a separate mapping S^ν in each iteration according to some rule, but anyway a lack of single-valuedness may come from several sources. In the case of numerical optimization, one source is often the use of inexact minimization in obtaining x_k from x_{k-1} through a subproblem such as line search. Again, the analysis of generalized continuity and differentiability properties of set-valued mappings is a critical need.

Integral Functionals and Measurability

Variational problems in an infinite-dimensional setting often involve expressions like

$$J_f(x) = \int_\Omega f\big(\omega, x(\omega)\big)d\omega \ \text{ for } \ x \in \mathcal{X}, \tag{7}$$

where Ω is perhaps a subset of $\mathbb{R}^d$ (but could also be a general measure space with measure $d\omega$, e.g. a probability space), and $\mathcal{X}$ is some space of functions $x : \Omega \to \mathbb{R}^n$ (perhaps a standard Banach space). The function f on $\Omega \times \mathbb{R}^n$ is an *integrand*, and J_f is an *integral functional*.

For example, a Lagrange problem of minimizing

$$\Phi(y) = \int_{t_0}^{t_1} f\big(t, y(t), \dot{y}(t)\big)dt$$

over all continuously differentiable functions $y : [t_0, t_1] \to \mathbb{R}^m$ satisfying $y(t_0) = a_0$, $y(t_1) = a_1$, with the notation $\dot{y} = dy/dt$, can be viewed in terms of $\Omega = [t_0, t_1]$ and $x(t) = \big(y(t), \dot{y}(t)\big)$. The problem is then seen as involving the composition of an integral functional J_f with a linear mapping $y \mapsto (y, \dot{y})$.

Classically, an integrand f in (7) would be assumed differentiable to whatever degree seemed convenient, but for modern purposes differentiability may not be taken for granted, because f could arise in some of the ways already suggested and be nonsmooth. Interestingly, just by allowing f to be extended-real-valued it is possible to achieve so much breadth, even in the otherwise ordinary statements of problems in the calculus of variations, that general problems of optimal control of ordinary and partial differential equations and variational inequalities are encompassed.

Of course, with the integrand f extended-real-valued one can hardly speak of continuity of f, much less differentiability, so in following this path the need for new methodology in the analysis of the functional J_f is inescapable. The very existence of J_f, in the sense of the integral being well defined for every function $x \in \mathcal{X}$, requires new foundations, because of technicalities concerning measurability of the function $\omega \mapsto f\big(\omega, x(\omega)\big)$ which cannot satisfactorily be dealt with in the old ways. A fresh approach has to be taken to the question of what it means for the function $f(\omega, \cdot\,)$ on $I\!R^n$ to depend measurably, as a whole, on the parameter element $\omega \in \Omega$. Very similarly, one needs to determine the right sense in which a set $S(\omega)$ depends measurably on ω.

In the case of a probability space, this is important in properly pinning down the notion of a *random* set or function. The relationship between random sets and random functions is seen in variational analysis through the fact that every extended-real-valued function on $I\!R^n$ can be identified with a specific subset of $I\!R^{n+1}$, namely its epigraph. On the other hand, an extended-real-valued function on $I\!R^n$ can be interpreted, in the manner explained above in connection with ∞ penalties, as fully designating a particular problem of optimization in n variables subject to constraints, so a random function could represent a random problem. Presumably, a random problem of optimization should give rise to a random optimal value and a random optimal solution set, but again, this terminology cannot legitimately be employed without first answering a host of technical questions about "measurability" that were never even posed in the classical setting. A sound theory of stochastic optimization is not possible without the elaboration of such details. Applications to variational principles in statistics have a stake in the matter too.

Measurable dependence on parameters is closely related to the generalized sorts of continuous dependence on parameters that have already been alluded to. This topic is finite-dimensional in a large way despite the integrand f and properties of the associated integral functional J_f on an infinite-dimensional space.

While the study of how sets and functions (through their epigraph sets) may depend measurably on parameters cannot properly be placed under the heading of variational analysis itself, it does furnish a strong example of how variational analysis interacts with other branches of mathematics and calls for major innovations in well established theories.

Generalized Differentiation

The notion of a variational inequality has already been described as providing a handle on the characterization of solutions to problems of optimization. In a larger scheme it is essential to develop generalized forms of differentiation that can be applied to a variety of situations lying beyond the capabilities of classical differential analysis.

On the geometric level, classical analysis can be said to revolve around the local study of *smooth manifolds*, sets which possess at each point a well defined *tangent space* and *normal space*, these being linear subspaces dual to each other in the sense of orthogonality. Such sets can in particular be the graphs of functions and vector-valued mappings, and in this way the geometry of tangents and normals may be translated into the analysis of directional derivatives, gradients, Jacobians and the like.

In variational analysis there is a similar pattern, but smooth manifolds are replaced by arbitrary closed sets, and instead of tangent and normal *spaces* one works with tangent and normal *cones*. A set is called a *cone* if it is nonempty and contains for each of its vectors w all multiples λw with $\lambda \geq 0$. Pictorially, a cone which is more than just $\{0\}$ is a bundle of rays (half-lines emanating from the origin). Subspaces are particular examples of cones.

For a closed set $C \subset I\!R^n$ and a point $\bar{x} \in C$, the *tangent cone* to C at $\bar{x}$, denoted by $T_C(\bar{x})$, consists of all the vectors w such that there exists a sequence of vectors $x_k \to \bar{x}$ in C along with a sequence of scalars $\lambda_k > 0$ such that $\lambda_k(x_k - \bar{x}) \to w$. On the other hand, the *normal cone* to C at $\bar{x}$, denoted by $N_C(\bar{x})$, consists of all the vectors v such that there exists a sequence of vectors $x_k \to \bar{x}$ in C along with a sequence of vectors $v_k \to v$ such that

$$\langle v_k,\, x - x_k\rangle \leq o(x - x_k) \;\text{ for }\; x \in C.$$

The cones $T_C(\bar{x})$ and $N_C(\bar{x})$ are always closed.

When C is a smooth manifold, $T_C(\bar{x})$ and $N_C(\bar{x})$ turn out to be the classical tangent and normal spaces. When C is a convex set, on the other hand, they are not subspaces orthogonal to each other but convex cones polar to each other, in the sense that

$$\begin{aligned} N_C(\bar{x}) &= \left\{\, v \,\middle|\, \langle v, w\rangle \leq 0 \;\text{ for all }\; w \in T_C(\bar{x}) \,\right\}, \\ T_C(\bar{x}) &= \left\{\, w \,\middle|\, \langle v, w\rangle \leq 0 \;\text{ for all }\; v \in N_C(\bar{x}) \,\right\}. \end{aligned} \tag{8}$$

In the convex case there is a simpler formula that can be used equivalently for the normal cone: one has

$$N_C(\bar{x}) = \left\{\, v \,\middle|\, \langle v,\, x - \bar{x}\rangle \leq 0 \;\text{ for all }\; x \in C \,\right\}.$$

Note that on the basis of this specialized formula of convex analysis the variational inequality (5) can be written as

$$-M(\bar{x}) \in N_C(\bar{x}).$$

The polarity relation (8) between tangent and normal cones holds not only when C is convex but in many other important cases as well, for instance when C is defined by a system of smooth constraints satisfying a constraint qualification as usually employed in the theory of optimization. Then C is said to be *Clarke regular* at $\bar{x}$. But the polarity relation does not hold universally. In the most general cases, the cones $T_C(\bar{x})$ and $N_C(\bar{x})$ can fail also to be convex. Nonetheless, they embody a great amount of local information about the set C which can be put to use effectively.

The lack of convexity and polarity just mentioned was viewed in the earlier stages of development of variational analysis as a serious drawback, and for this reason the definitions of $T_C(\bar{x})$ and $N_C(\bar{x})$ given here were not the ones used. Alternative definitions of the tangent and normal cones, which did yield convex cones, were employed instead; specifically the normal cone was taken to be the convex hull of the set denoted here by $N_C(\bar{x})$, and the tangent cone was taken to be its polar. This is the pattern followed by Clarke in [3] and [4], for instance, as well as in other works by a multitude of authors, including the present writer. Through the work of Mordukhovich [5], however, it has come to light that such convexification can be bypassed almost entirely (except in some special circumstances where integral functionals are involved), and that the results obtained then are even stronger. An updated view of the subject is being put together in [6].

The geometry of tangents and normals can be translated into concepts of analysis through consideration of epigraphs. Consider a general function $f : I\!R^n \to \overline{I\!R}$ and a point $\bar{x}$ where $f(\bar{x})$ is finite. The epigraph

$$E = \left\{ (x, \alpha) \in I\!R^n \times I\!R \,\middle|\, \alpha \geq f(x) \right\}$$

is a closed set in $I\!R^{n+1}$ with the point $\big(\bar{x}, f(\bar{x})\big)$ on its boundary. The cones $T_E\big(\bar{x}, f(\bar{x})\big)i$ and $N_E\big(\bar{x}, f(\bar{x})\big)$ provide the needed handle on the local properties of generalized differentiability of f at $\bar{x}$.

The *subderivative* function associated with f at $\bar{x}$ is the function $df(\bar{x}) : I\!R^n \to \overline{I\!R}$ defined by

$$df(\bar{x}) = [\text{ function having the set } T_E\big(\bar{x}, f(\bar{x})\big) \text{ as its epigraph }],$$

while the *subgradient* set is the set $\partial f(\bar{x}) \subset I\!R^n$ defined by

$$\partial f(\bar{x}) = \left\{ v \in I\!R^n \,\middle|\, (v, -1) \in N_E\big(\bar{x}, f(\bar{x})\big) \right\}.$$

In general, the subderivative function can always be expressed by the alternative formula

$$df(\bar{x})(\bar{w}) = \liminf_{\substack{w\to\bar{w}\\ t\searrow 0}} \big[f(\bar{x}+tw) - f(\bar{x})\big]/t.$$

On the other hand, it can be shown that $v \in \partial f(\bar{x})$ if and only if there is a sequence of points $x_k \to \bar{x}$ with $f(x_k) \to f(\bar{x})$ along with a sequence of vectors $v_k \to v$ such that

$$f(x) \geq f(x_k) + \langle v_k,\, x - x_k\rangle + o(x - x_k).$$

In many cases, however, simpler formulas can be substituted which give the same result.

For instance, when f is convex, one actually has

$$\partial f(\bar{x}) = \big\{\, v \in I\!R^n \,\big|\, f(x) \geq f(\bar{x}) + \langle v,\, x-\bar{x}\rangle \text{ for all } x \,\big\}.$$

When f is not only convex but finite on a neighborhood of $\bar{x}$, one further has

$$df(\bar{x})(\bar{w}) = \lim_{t\searrow 0} \big[f(\bar{x}+t\bar{w}) - f(\bar{x})\big]/t.$$

The polarity in (8) is then available and translates into the relation

$$df(\bar{x})(\bar{w}) = \max_{v\in\partial f(\bar{x})} \langle v, \bar{w}\rangle. \tag{9}$$

Equation (9), which also holds in many other cases beyond the convex one (although not always), neatly generalizes the classical equation for directional derivatives of f in terms of the gradient $\nabla f(\bar{x})$ when f is differentiable at $\bar{x}$:

$$df(\bar{x})(\bar{w}) = \langle \nabla f(\bar{x}), \bar{w}\rangle.$$

In fact, the latter is obtained from (9) when the set $\partial f(\bar{x})$ consists of just a single vector v. For a convex function, or a Lipschitz continuous function, such is the case if and only if f differentiable at $\bar{x}$, the unique subgradient v then being $\nabla f(\bar{x})$.

The elements of $\partial f(\bar{x})$ are called the *subgradients* of f at $\bar{x}$. They can be used to express optimality conditions. An elementary first-order condition for $\bar{x}$ to be optimal for the problem of minimizing f is

$$0 \in \partial f(\bar{x}).$$

This relation obviously generalizes Fermat's rule $0 = \nabla f(\bar{x})$ for the case where f is differentiable. It would be of little value in itself, were it not for the existence of a robust machinery for the *calculus* of subgradients. For instance, under a mild assumption corresponding to a constraint qualification, one has

$$\partial(f_1 + f_2)(\bar{x}) = \partial f_1(\bar{x}) + f_2(\bar{x})$$

in the sense that the elements v of the set on the left are precisely the vectors of the form $v_1 + v_2$ with v_1 and v_2 selected from the sets on the right. Interestingly too, as a tie-in with the monotone mappings mentioned earlier, if f is convex, then the set-valued mapping $\partial f : x \mapsto \partial f(x)$ is maximal monotone.

Generalized higher derivatives of nonsmooth functions are also beginning to be understood. Broad second-order conditions for optimality are now available, for example. For this topic we refer to [7], [8] and [9].

References

1. R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
2. R. T. Rockafellar, *Conjugate Duality and Optimization*, SIAM Publications, 1974.
3. F. H. Clarke, *Nonsmooth Analysis and Optimization*, Wiley, 1983; now republished by SIAM Publications, 1991.
4. F. H. Clarke, *Methods of Dynamics and Nonsmooth Optimization*, SIAM Publications, 1989.
5. B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, 1988 (Russian); about to appear in English translation (Wiley).
6. R. T. Rockafellar and R. J-B Wets, *Variational Analysis*, Springer-Verlag sometime; not yet finished!
7. R. T. Rockafellar, "First and second-order epi-differentiability in nonlinear programming," Transactions Amer. Math. Soc. 307 (1988), 75–108.
8. R. T. Rockafellar, "Second-order optimality conditions in nonlinear programming obtained by way of epiderivatives," Math. of Op. Res. 14 (1989), 462–484.
9. R. T. Rockafellar, "Generalized second derivatives of convex functions and saddle functions," Transactions Amer. Math. Soc. 320 (1990), 810–822.