Journal of Computational and Applied Mathematics 22 (1988) 243–255
North-Holland

Optimization: a case for the development of new mathematical concepts

R.T. ROCKAFELLAR
Department of Mathematics, University of Washington, Seattle, WA 98195, U.S.A.

Received 1 April 1987

Keywords: Optimization, linear programming, mathematical modeling, duality, games, nonsmooth analysis.

Optimization is a big subject with deep historical roots, but its scope and form are radically different today from what they one were. A revolution that begin in the 1950s has changed the entire outlook and generated a new kind of mathematics with far-reaching consequences, not only for applications but for the very foundations of analysis.

It might be imagined that this revolution is attributable to the advent of the computer, and to some extent this is true. Armed with the capabilities of computers, it became possible to attack problems that previously were beyond solution. To think, though, that the revolution mainly had to do with finding new ways of processing input and output from computers in order to bridge a gap between theory and practice, would be a mistake. Rather it had to do with the different nature of the emerging problems and the inadequacy of traditional mathematical concepts in dealing with them.

Mathematics, after all, is a living body of thought which has grown through the centuries in response to the challenges of various civilizations and periods. This in not to say that mathematics only progresses out of specific need, because it clearly also has its own internal lines of development. But the geometry of the Greeks would not have arrived without the preoccupation of the ancient world with architecture and public works. The tone of much of the mathematics of present times was set by the triumph of the physical sciences in the eighteenth and nineteenth centuries. Almost everything about differential equations falls into this category, for example. A fundamentally modern innovation, the theory of probability and statistics, receives impetus from contemporary pressure in the social sciences, medicine and agriculture.

The purpose of this article is to explain briefly the nature of optimization as a rapidly expanding field and the importance of fundamental mathematical research in its continued success. The main point is that the applicability and computational vigor of any subject depend heavily on the conceptual framework that is available. This framework needs to be developed with originality, rigor and an eye for the right level of abstraction. Such development will not take place if left in the hands of practitioners attending to just a narrow range of applications. It requires the dedicated efforts of trained mathematicians.

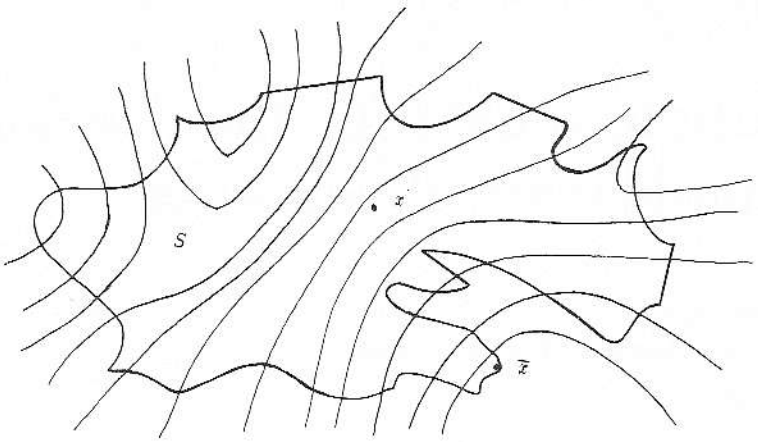

Fig. 1.

The fundamentals of optimization

The typical circumstance that gives rise to a problem of optimization is a need for comparing or deciding among the elements of some set S. These elements may represent the possible states of a system or the available choices in some situation. Often they can be specified by giving the values of finitely many variables $x_1, \ldots, x_n$, and in this way identified with points $x = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$. Then S is a subset of $\mathbb{R}^n$, but usually not all of $\mathbb{R}^n$: the points x in S must satisfy certain equations or inequalities, called *constraints*. In applications where the elements of S cannot be so described, perhaps because they represent functions or probability distributions, S may not be a subset of $\mathbb{R}^n$, of course, but of some infinite-dimensional space.

For the sake of comparing or deciding among the elements of S, there must be some criterion. Ordinarily this is given by a real-valued function f, called the *objective* function. The basic problem of optimization is

$$\text{minimize } f(x) \text{ over all } x \text{ in } S.$$

Maximization could be the goal rather than minimization, but for theoretical purposes there is no difference, because maximizing a function g is equivalent to minimizing $f = -g$.

The elements x of S are called the *feasible solutions* to the problem, while an $\bar{x}$ giving the minimum is called an *optimal solution*. (See also Fig. 1.) The number $f(\bar{x})$ is the *optimal value*. In specific applications it is sometimes difficult even to determine a feasible solution, because of complications in the constraints that define S. Sometimes it is just the optimal value that is derived, but in other cases the function f is merely an artificial construct and only an optimal solution is acceptable. Always, however, the fact that an optimal solution may occur along the boundary of S at a potentially 'nasty' point is a source of trouble.

The calculation of optimal solutions and values is, in the long run, a primary concern, but a great amount of study is needed before the stage can be reached where calculations can effectively be performed and the results fully interpreted. Necessary and sufficient conditions for optimality must be discovered, not only for use as tests in algorithms, but as a guide for understanding the situation that is being modeled. Questions about the sensitivity or stability of

the problem to perturbations in the structure of S and f must be answered. Problem types must be classified according to their amenability to various approaches. Different formulations must be compared for their advantages and disadvantages, and the methods of passing between them clarified. Criteria must be set up for recognizing whether the properties on which a conclusion or numerical procedure may depend are present or not.

All this can have a very important influence on the modelling process itself. Mathematical modeling in the context of any sort of application is an art that obviously depends on knowing what models are available and how effectively they can be handled. There is no such thing as simply using a computer to get the answers one needs. Any numbers that are computed have relevance only to the model that has been selected, not to speak of the computational method. This model may be convenient or inconvenient, too coarse or too fine, and the conclusions based on it may be mathematically correct or incorrect.

The truth of the matter is that classical analysis does not furnish suitable models for most of the modern situations involving optimization, because the problems are often in areas like economics or management that previous generations of mathematicians did not think much about. New approaches have had to be invented. Optimization theory serves therefore as an excellent example of mathematical ideas at work, where there is a vital connection between theoretical progress and rapidly expanding applications. This is much to the consternation of oldtimers, incidentally, who think they know how to tell the difference between what is 'pure' and what is 'applied'. Such may be possible in a branch of mathematics that is cut and dried, although one could dispute it. Certainly it is not true in the subject we are going to look at here.

The range of applications

Elementary problems of optimization are encountered by every student in first-year calculus. What proportions in a cardboard box with an open top will minimize the amount of cardboard needed to achieve a given volume? What proportions in a metal can will maximize the volume for a given amount of metal? These are questions in engineering design.

More complicated problems of engineering design are easy to find. What shape of an aircraft wing minimizes drag, subject to maintaining the desired flight characteristics? What structure of a building minimizes cost subject to staying earthquake-proof? How should the components of an electronic unit be arranged in order to minimize signal loss or interference?

Other problems of optimization are concerned not so much with design as with the management of systems already in place. What trajectory should a satellite launching rocket follow in order to achieve orbit with the least amount of fuel? What mixture of available ores provides the cheapest way of manufacturing a given alloy to within specifications? What schedule of drawdowns should be followed for a network of reservoirs to ensure the most reliable level of hydroelectric generating capacity without jeopardizing essential irrigation usage? How should a pattern of insecticide spraying be organized in order to wipe out an infestation at least damage to the environment?

Often the concerns are distinctly economic in nature. How should production and inventories be managed over time in order to maximize expected profit in the face of uncertain demand? How should capital investment be carried out in order to attain the highest rate of economic growth within a given period of time?

Optimization enters many applications in a purely mathematical way, too. Typical of this are problems of numerical approximation or the identification of parameters. In the first case one tries to find the simple function from a restricted class that best fits a given, more complicated function. In the second case a particular model has been chosen for some phenomenon, but the parameters in the model have yet to be assigned their specific values. The parameters must give the best fit in some sense to a mass of empirical data. A kind of 'distance' expression must be minimized. Nonlinear regression and maximum likelihood estimation are examples of parameter identification problems in statistics that have both theoretical and practical components.

Finally there are important problems of optimization that do not involve decision-making at all but relate rather to characterizing the behavior of a system that operates autonomously. In this category are the classical variational principles of physics and chemistry: a system minimizes "action", or tends toward a state of minimal energy. The recognition that the equilibrium states in such a setting may be viewed as optimal solutions to problems of constrained minimization can have important consequences in science. The equilibrium mixture of gases in a planetary atmosphere, or of biochemical species in human blood, can be determined despite the large number of reactions possible within the system and across its boundaries by minimizing the Gibbs free energy subject to a set of mass balance conditions. On the other hand, one may now be able to extend the basic theory of various physical systems to nonclassical situations involving multiple unilateral constraints, for instance, by invoking the pertinent variational principle and passing to the optimality conditions that the case entails.

A large part of mathematical economics is likewise in this category of optimization. It is concerned with the identification and analysis of variational principles, such as maximization of 'utility', that govern economic behavior. It provides many interesting questions which push on the limits of present theory. Multiple objectives must be considered, and also conflicts between different agents or decision makers. The theory of many-person games is a fascinating example of optimization in this mode.

Departure from classical notions

In the framework for optimization provided by classical analysis, the function f and set S are comparatively simple in structure. There are only a few constraints, and these are usually in the form of *equations*. Differentiability is taken for granted. Thus in finite-dimensional problems S is a curve, surface, or other smooth manifold, or at worst a closed region of such a manifold whose boundary is composed of a small number of nice pieces that can readily be described and given a parametric representation; see Fig. 2(a).

There is a symmetry in attitude between minimization and maximization. Indeed, one is almost as interested in 'stationary points' as in true extrema. The task is viewed in terms of determining all possible stationary points of f relative to S and classifying them by their nature, whether they give minima, maxima, or something in between. For this an entirely local analysis involving first and second derivatives is deemed adequate.

Modern problems fail to fit this framework for a variety of reasons. First of all, the number of constraints can be very large, and these constraints are chiefly in the form of inequalities rather than equations. This means that the geometry of the set S can be very complicated. The equations among these constraints may still determine a smooth manifold in which S lies, but

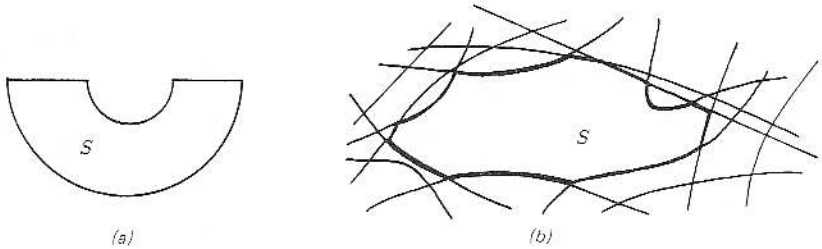

Fig. 2.

the boundary of S as a close region in that manifold can be made up of an astronomical number of pieces which do not fit together in any easily describable pattern; see Fig. 2(b). It no longer is reasonable to think of analyzing the boundary one piece at a time, each piece having an individual representation. It may be preferable to view the boundary rather as a sort of 'nonsmooth hypersurface'. The graph of the objective function f in many applications turns out to have such a character too. This is strong motivation for the development of new mathematical techniques which resemble the calculus but do not depend on the existence of derivatives and gradients in the classical sense.

Another distinguishing feature of modern problems is an unambiguous orientation towards either minimization or maximization. Hardly ever is one interested in both in the same context, i.e. with respect to the same set of variables, much less in the question of more general stationary points, except as potential troublespots in the convergence behavior of optimization algorithms. Furthermore, a *global* analysis of optimality is desired.

Modern problems are also typically much larger in scale than was true in the past. There is no hope of solving them by hand and therefore little reason to pursue the kinds of 'closed form' methods of solution that used to be the norm. Problem structure must be studied instead with a view towards convenience of representation and manipulation in a computer. Ideas of decomposition or the decentralization of decision making become very attractive, and not only for purposes of computation. These ideas are powerful tools in the modeling phase too.

Stochastic elements are yet another complication. Problems may involve random variables and conditional expectations. They may suffer from uncertainties in the data and require an adaptive approach which involves the selective generation of fresh data. Monte Carlo techniques may offer a way of getting around local obstacles in order to discover a globally optimal solution. Such probabilistic considerations can carry the analysis far from the classical frame of reference.

The example of linear programming

Linear programming problems were among the first in optimization to break the old bounds. They came into focus around 1950, just when computers were getting going and the idea had dawned that mathematics could be applied to questions of logistics, production and management. 'Programming' was originally just a bureaucratic synonym for 'planning', as in setting up a government program to accomplish a certain task. When computers were involved, one had 'computer programming'. When optimization theory was involved, one had 'mathematical programming'. Nowadays programming has come to mean giving instructions to a computer, at

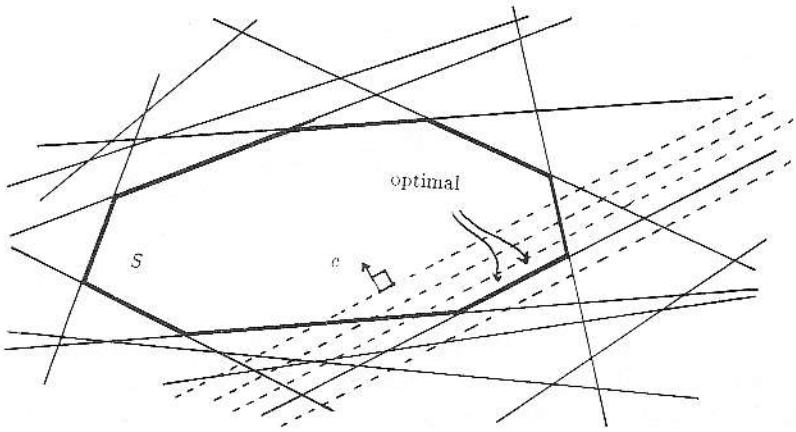

Fig. 3.

least to most people. But in mathematics a different usage has evolved: linear and nonlinear programming, convex programming, integer programming, and network programming all refer to areas of optimization which are distinguished from each other by the character of objectives and constraints.

In linear programming the objective function is linear and the constraints are in general a mixture of linear equations and weak linear inequalities. Tricks of reformulation, however, make it possible to cast any linear programming problem in the following standardized form:

$$\begin{aligned} &\text{minimize} && c_1x_1 + c_2x_2 + \cdots + c_nx_n, \\ &\text{subject to} && a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \geqslant b_1, \\ & && a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \geqslant b_2, \\ & && \vdots \\ & && a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \geqslant b_m. \end{aligned} \tag{P}$$

Each of the constraints defines a closed half-space in the n-dimensional space $\mathbb{R}^n$ of points $x = (x_1, \ldots, x_n)$. The feasible solution set S is the intersection of these half-spaces, as in Fig. 3. A set S of this kind is called a *convex polyhedron*.

Systems of linear equations have been an object of study for a very long time, of course. How remarkable it is, therefore, and what a comment on the way mathematics proceeds, that when linear programming came into being there was barely any theory of linear inequalities on hand. Hardly anyone had thought about the subject before or imagined it to be worth developing. The traditional mind-set was strong, and many people actually found it hard to believe that constraints could often be formulated more appropriately as inequalities than equations. To this day one can see papers in engineering or economics, for instance, that fail to appreciate this point.

But if the inequalities in problem (P) were written as equations, this could be a severe restriction that might even preclude the existence of any feasible solution at all, not to speak of an optimal solution. Suppose for example that (P) is a blending problem where n different substances are to combined in the cheapest possible way that meets certain minimal require-

ments on the final composition. Then x_j denotes the amount of substance j to be added to the mixture, and c_j is the cost per unit of that substance, so the objective expression gives total cost. The minimal requirement for ingredient i in the mixture is b_i, and a_{ij} is the amount of this ingredient per unit of substance j. Each constraint then represents a condition that the total amount of an ingredient i present in the mixture must be *at least* as high as the minimum required. To insist that it actually *equal* the minimal amount for every i would be to exclude various potentially cheaper mixtures from consideration and even to run the risk of there being no mixture whatsoever that meets such tight conditions, regardless of cost.

The linear programming problem (P) illustrates another feature that is blithely accepted in optimization today but was anathema in the past: a possible multiplicity of solutions. Inasmuch as the objective function is linear, its level contours (where a constant value is assumed) form a family of parallel hyperplanes (dotted lines in Fig. 3). If these happen to line up with one of the 'faces' of the convex polyhedron S, then that entire face will consist of optimal solutions. This may seem like a rare event, since the slightest perturbation of coefficients would preclude it. In practice, however, it occurs quite often, because the constraints of a problem, far from being 'random', are usually in special relationships with the objective and each other.

Everyone has therefore gotten accustomed to speaking of *an* optimal solution rather than *the* optimal solution. It is interesting to contrast that with the traditional thinking that a mathematical problem can hardly be considered well posed unless the uniqueness of the solution, at least locally, can be established along side of the existence. That frame of mind is due, of course, to past reliance on equations as the work horse in mathematical modeling. In a rough sense it is true that, with equations, nonuniqueness can occur only with some kind of mismatch between the number of conditions that have been identified and the number of degrees of freedom that are present. But this has no validity when inequalities are the rule.

Although a linear programming problem may have more than one optimal solution, there will always be among them at least one that is a *vertex* of the convex polyhedron S. It follows that in order to be solve such a problem, all one has to do is inspect the vertices one by one (there being only finitely many) along with the corresponding values of the objective function. A vertex yielding the lowest such value is an optimal solution to (P). In algebraic terms, vertices are characterized as the intersections of various collections of the hyperplanes that bound the constraining halfspaces. They can therefore be determined by solving various systems of equations generated by the inequalities in (P).

Here again we have an illustration of how attitudes have changed. Back in the 1950s, many mathematicians when presented with this description of solving a linear programming problem would have felt there was little left to be said. They might well have felt let down: the problem had turned out to be 'trivial'. To be sure, it might not be an easy matter to inspect all the vertices when there are many of them, but that is just a question of tedium. A computer could do it, some day anyway. From the mathematical standpoint, once a problem had been reduced to the enumeration of finitely many possibilities, nothing of interest could remain.

To say there are 'many' vertices, however, is comically mild. Linear programming problems that today can routinely be solved in seconds may have more vertices than there are grains of sand on all the planets in the universe! How then is their solution possible? It is possible because the set of vertices has significant structure and should not be viewed merely as discrete. Two vertices of S can be adjacent to each other; one can pass between them by following an 'edge' of S. By starting at one vertex and moving along edges, always in the direction of improvement in

the value of the objective, one can reach an optimal vertex in relatively short time. Optimality is detected by the lack of any adjoining edge along which a further improvement is possible. This, in essence, is the celebrated Simplex Method of G.B. Dantzig, which for its practical applications in the last 35 years must be regarded as one of the most fruitful discoveries in modern mathematics.

Needless to say, the Simplex Method is not as elementary to execute as it sounds. A considerable theoretical back-up is required in tying algebra and geometry together, devising tests for feasibility and optimality of vertices, avoiding possible 'degeneracy', and establishing the typical amount of time that the method may be expected to take. This all has had to be worked out by mathematicians in theorem-and-proof style.

Duality and other surprises

Linear programming theory has many other novel characteristics besides the ones already mentioned. Among the most striking is the phenomenon of duality, which had no recognized precedent in classical optimization but is now seen as fundamental to many kinds of problems.

One way of arriving at duality is through the conditions that identify a feasible solution to a linear programming problem (P) as optimal. The basic picture is displayed in Fig. 4. At the point $\bar{x}$ of S only some of the inequality constraints are 'active', i.e. satisfied with equality; the others are 'slack', i.e. satisfied with strict inequality. The gradients of the active constraints at $\bar{x}$ are certain vectors $a_i = (a_{i1}, \ldots, a_{in})$, whereas the gradient of the linear objective f is the vector $c = (c_1, \ldots, c_n)$. The condition which is necessary and sufficient for optimality is following: c should belong to the 'cone' K generated by taking all linear combinations of the active constraint gradient a_i with nonnegative coefficients.

It is convenient in expressing this property to assign coefficients also to the inactive constraint gradients but require them to be zero. The optimality condition on $\bar{x}$ then amounts to the existence of a coefficient vector $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_m)$ which together with $\bar{x}$ satisfies

$$\bar{y}_i \geqslant 0, \quad a_i \bar{y}_i - b_i = 0, \quad \bar{y}_i (a_i - \bar{x} - b_i) = 0,$$
$$\bar{y}_1 a_1 + \bar{y}_2 a_2 + \cdots + \bar{y}_m a_m = c. \tag{C}$$

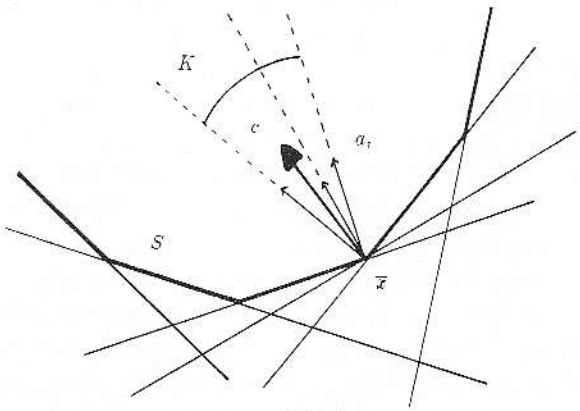

Fig. 4.

Here the first row expresses the feasibility of $\bar{x}$ and the requirement that for each i with $a_i \cdot \bar{x} - b_i < 0$, one must have $\bar{y}_i = 0$. Of course, $a_i \cdot \bar{x} = a_{i1}\bar{x}_1 + \cdots + a_{in}\bar{x}_n$.

There is nothing particularly remarkable so far in this. True, the process of finding an optimal solution $(\bar{x}_1, \ldots, \bar{x}_n)$ to (P) appears to involve the determination of a special coefficient vector $(\bar{y}_1, \ldots, \bar{y}_m)$. That sort of thing occurs even classically, however; the coefficients $\bar{y}_i$ are evidently to be regarded as Lagrange multipliers associated with the constraints in (P).

A surprise comes, though, with the realization that these coefficients are connected with a second linear programming problem, namely

$$\begin{aligned} &\text{maximize} && y_1 b_1 + y_2 b_2 + \ldots + y_m b_m, \\ &\text{subject to} && y_1 a_{11} + y_2 a_{21} + \ldots + y_m a_{m1} = c_1, \\ & && y_1 a_{12} + y_2 a_{22} + \ldots + y_m a_{m2} = c_2, \\ & && \vdots \\ & && y_1 a_{1m} + y_2 a_{2m} + \ldots + y_m a_{mn} = c_n, \\ & && y_1 \geqslant 0, \quad y_2 \geqslant 0, \quad \ldots, \quad y_m \geqslant 0. \end{aligned} \tag{D}$$

This is called the *dual* of problem (P). The crucial fact is that when we analyze what it means for a feasible solution $\bar{y}$ to (D) to be optimal, as we did in the case of (P), the necessary and sufficient condition we get is the following: there should exist a vector $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ such that (C) is satisfied—the same (C) as before!

Optimality in (P) and (D) thus receives a joint characterization. The $\bar{y}_i$'s may be Lagrange multipliers for the constraints in (P), but at the same time the $\bar{x}_j$'s are Lagrange multipliers in for the constraints in (D). Moreover the optimal values of the objectives in the two problems turn out to be tied together by this: one has

$$\min \text{ in (P)} = \max \text{ in (D)}.$$

In brief, neither problem (P) nor problem (D) can be solved without automatically solving the other. Indeed, the Simplex Method does solve both and actually depends on this property in its algebraic formulation.

The more one thinks about this, the more mysterious and intriguing it seems. We thought we had only one optimization problem. Why couldn't matters stay that simple? What interpretation can the hidden second problem possibly have, especially in concrete applications?

Even more surprising than the phenomenon of duality itself is the nature of the key to its explanation, namely the theory of games. This theory, which is closely related to optimization in more ways than one, was invented by J. von Neumann, one of the greatest mathematicians of the present century. (Besides games and other landmark contributions to mathematical economics, he worked on the mathematical foundations of quantum mechanics and, before his untimely death in the early 50s, was instrumental in the development of computers).

In a so-called two-person zero-sum game, there are two players, I and II, each of which has a set of elements called 'strategies'. Player I selects an element x from his set U, while Player II selects an element y from his set V. They make the selections secretly and reveal their choices simultaneously. Then there is a 'pay off' of an amount $L(x, y)$ (positive, negative, or zero) from Player I to Player II.

Obviously Player I would like to minimize the pay-off amount while Player II would like to maximize it, but neither player has complete control over the outcome. Nevertheless it is possible

to set up rational criteria for play, on the basis of which Player I chooses an optimal element $\bar{x}$ by solving a certain minimization problem which is independent of the unknown action to be taken by Player II, and similarly Player II chooses an optimal element $\bar{y}$ by solving a certain maximization problem which is independent of the unknown action to be taken by Player I. In this manner any two-person zero-sum game gives rise to a pair of simple optimization problems, one of minimization and the other of maximization. Under certain fairly general assumptions, it can be established that the minimum value in the one problem equals the maximum value in the other. The optimal elements $\bar{x}$ and $\bar{y}$ then furnish a sort of state of equilibrium, and intuitively it is clear why in such cases the optimality of $\bar{x}$ and $\bar{y}$ might have a joint characterization.

The notion of a two-person zero-sum game attempts to provide an abstract mathematical model for the most fundamental instance of conflict in decision making. This model may seem too simple to be of much use, but one can demonstrate that it actually covers games like chess and poker as well as numerous situations in economics and human affairs. Our purpose in mentioning it here, however, is its role in the theory of linear programming.

It turns out that the linear programming problems (P) and (D) can be identified with the separate strategy problems for Players I and II in a certain two-person zero-sum game. In this game, the set U from which Player I makes his selection is the set of all vectors $x = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$, while the set V for Player II consists of all the *nonnegative* vectors $y = (y_1, \ldots, y_m)$ in $\mathbb{R}^m$. The 'pay-off' $L(x, y)$ for x in U and y in V is given by

$$L(x, y) = c_1x_1 + \cdots + c_nx_n + b_1y_1 + \cdots + b_my_m - a_{11}y_1x_1 - a_{12}y_1x_2 - \cdots - a_{mn}y_mx_n.$$

By analyzing the meaning of this function in a particular application, one can arrive at an interpretation of the dual problem and the information furnished by its solution $\bar{y}$. Since, after all, this vector of coefficients is automatically going to be produced by any method that solves the given problem (P), there is good reason for trying to understand its significance and the possible uses to which it might be put.

The game-theoretic interpretation of linear programming duality provides many answers, but it does not really dispel the mystery. It tells us that in trying to solve a linear programming problem, innocent as that may seem, we become party to a sort of metaphysical conflict. There is an antagonist out there whose interests are directly opposed to ours. Remarkably, this happens not only in linear programming but in many other areas of optimization as well. A great amount of mathematical effort has gone into exploring this notion and marking out its limits, but the job is far from done yet. What better example can there be of the fresh and meaningful challenges that continually arise in mathematics?

Convexity and nonsmooth analysis

The complicated nature of the boundary of the feasible set S in an optimization problem gives incentive for developing a new kind of 'nonsmooth' calculus, as mentioned earlier. There are other incentives too, and these can in part be understood after further examination of the linear programming pair (P) and (D).

These problems depend on the specification of coefficients a_{ij}, b_i and c_j for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, but for the time being let us think of the a_{ij}'s and c_j's as fixed and the b_i's as

parameters that can vary. The common optimal value in the two problems is then a function of the vector $b = (b_1, \ldots, b_m)$:

$$v(b) = \min(P) = \max(D).$$

The optimal solutions to the two problems are also in some sense functions of b, but the catch here is that the optimal solutions are not necessarily unique. Therefore one cannot speak of a function in strict terms, which would require unambiguous single-valuedness. The word *multifunction* has come into use for something that assigns to each point not always a single value but possible a set of values (maybe the empty set in some cases). Thus we have two multifunctions X and Y here rather than functions:

$X(b)$ = set of all optimal solutions to (P),

$Y(b)$ = set of all optimal solutions to (D).

Of course for many, even most choices of b these sets may reduce to single elements.

It goes without saying, that great interest resides in the question of how $v(b)$, $X(b)$ and $Y(b)$ behave when b is perturbed. What kind of continuity properties, if any, can be expected? Is it possible to quantify the changes in terms of 'rates'?

As far as the multifunctions X and Y are concerned, it is immediately clear that new ideas are needed beyond anything available in standard calculus. If $X(b)$ can be a set rather than a single point, what sense should one ascribe to $X(b')$ converging to $X(b)$ as b' converges to b, and even if an intuitive idea can be formulated, does this property typify what actually happens in the case of an optimal solution multifunction? How can one pretend to form a difference quotient $[X(b+th) - X(b)]/t$ in terms of sets, and is this the appropriate thing to try to do anyway?

The function v may at first seem easier to deal with, but serious difficulties lurk under the surface. The critical feature is that the very definition of v involves a process of optimization in several variables, whether we view it in terms of (P) or (D). Standard calculus is not designed to treat functions constructed in such a manner. It can handle arithmetical operations like addition and multiplication, and also the composition of one function with another, and formulas involving integration, but formulas involves minimization or maximization are beyond its domain. The inescapable fact is that functions defined by such formulas generally fail to be differentiable in the usual sense. The inherited theorems about derivatives and gradients therefore cannot be applied.

Some inkling of the state of affairs can be obtained by exploring the nature of the optimal value function v at a point b where the dual optimal solution set $Y(b)$ happens to reduce to a single vector $\bar{y}$. From an analysis of the optimality conditions (C) it is possible to show that for all b' in some region around b, $Y(b')$ likewise reduces to $\bar{y}$, and

$$v(b') = v(b) + \bar{y} \cdot (b' - b).$$

Thus v is linear around b with gradient $\nabla v(b) = \bar{y}$ This hints at a close connection between the differential properties of v and the optimal solutions to (D), but the same $\bar{y}$ cannot be the unique optimal solution to (D) for every choice of b, so v cannot be a linear function in the large. What is the overall picture of v, and how does it account for the fact that $Y(b)$ can sometimes contain more than one element?

It turns out that v is 'piecewise linear', but not just in an arbitrary way; see Fig. 5. The graph of v is the lower boundary of a certain convex polyhedron. The points b where $Y(b)$ reduces to a

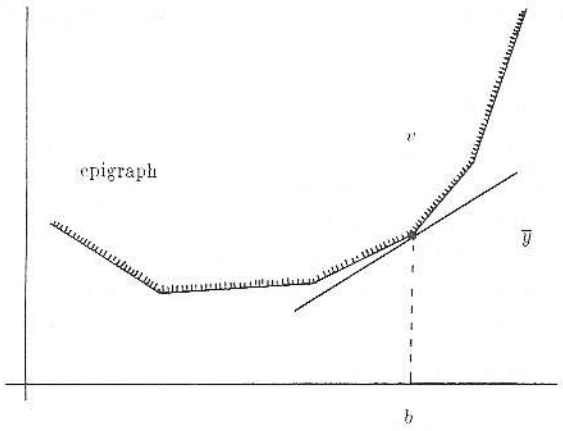

Fig. 5.

single $\bar{y}$ correspond to the flat faces of the polyhedron. At points b that correspond to edges or corners of the polyhedron, there is no uniquely determined 'linearization' of v but just a set of possibilities, a set of vectors $\bar{y}$. This set happens exactly to be $Y(b)$!

The connection between v and Y is therefore very close indeed. If we are willing to consider *one-sided* directional derivatives as defined by

$$v'(b;\ h) = \lim_{t \downarrow 0} \frac{v(b+th) - v(b)}{t}$$

we can come to the conclusion that for each b this expression, considered as a function of h, is well defined and completely characterized by the set $Y(b)$, and vice versa:

$$v'(b;\ h) = \max_{\bar{y} \in Y(b)} \bar{y} \cdot h \quad \text{for all } h.$$

This formula is powerful testimony to the role of the dual problem (D) in the behavior of the optimal value in problem (P). When $Y(b)$ reduces to a single $\bar{y}$, the formula turns into the gradient relation

$$v'(b;\ h) = \bar{y} \cdot h = \nabla v(b) \cdot h.$$

Several lessons can be drawn from this example. By a new approach, the concepts of calculus can be adapted and extended into a vast new territory. The classical geometric emphasis on the graph of a function can be replaced by an emphasis on *epigraph*, which consists of all the points lying on *or above* the graph. Such an approach may seem unsymmetric, but then a fundamental lack of symmetry must already be recognized at the heart of optimization theory.

Two-sided derivatives can be replaced by various one-sided limits. Instead of a 'gradient', one can look towards a characterization of the generalized derivatives by a set of 'subgradients'. Thus for the function v one obtains by the process of differentiation not a gradient function ∇v, but a subgradient *multifunction* ∂v, given by $\partial v(b) = Y(b)$. Multivaluedness, then, is something that makes sense on more than one level and can be welcomed as a natural expression of reality.

Another important idea is that of using convexity as a tool where linearity and linearization no longer work. This has proved to be highly successful not only in treating functions whose epigraphs are convex polyhedra, as in Fig. 5, but for extremely general classes of nonsmooth

functions whose epigraphs are in no way convex. Such functions arise as optimal value functions v for more complicated kinds of optimization problems than (P). Their study is yielding valuable information about stability and methods of decentralization or decomposition. The new forms of analysis are also being applied now to infinite-dimensional problems and variational principles, and in this way they are affecting more traditional areas of mathematics, like the theory of partial differential equations.

All of this leads to the conclusion that mathematics, as the science of identifying the essential geometrical and numerical concepts in any situation and making the most of them, is as vital to progress today as in past eras of discovery. The arrival of computers, far from signaling that everything is over except for the computations, has shaken long-standing ideas and presented new challenges that mathematicians will long be working hard to meet.

References

[1] F. H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley-Interscience, New York, 1983).
[2] R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, NJ, 1970).
[3] G.B. Dantzig, *Linear Programming and Its Extensions* (Princeton University Press, Princeton, NJ, 1963).
[4] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ, 1953).