CONCEPTS OF OPTIMALITY AND THEIR USES

Nobel Memorial Lecture, December 11, 1975

by
TJALLING C. KOOPMANS
Yale University, New Haven, Connecticut, USA

According to a frequently cited definition, economics is the study of "best use of scarce resources." The definition is incomplete. "Second best" use of resources, and outright wasteful uses, have equal claim to attention. They are the other side of the coin.

For our present purpose the phrase "best use of scarce resources" will suffice. However, each of the two nouns and two adjectives in this phrase needs further definition. These definitions in turn need to be varied and adjusted to fit the specific circumstances in which the various kinds of optimizing economic decisions are to be taken.

I will assume that the main interest of this gathering is in the range of applications of the idea of best use of scarce resources, and in the ways in which the main categories of applications differ from each other. I shall therefore describe mathematical ideas and techniques only to an extent helpful for the exploration of that range of applications.

A good place to start is with the *production programs of the individual plant or enterprise* for a short period ahead. The "resources" then include the capacities of the various available pieces of equipment. In a centrally directed economy they may also include the allotments of nationally allocated primary inputs such as fuels, raw materials, labor services. In a market economy with some capital rationing one single allotment of working capital available for the purchase of primary inputs at given market prices would take the place of most of the primary input allotments.

In either institutional framework, an especially simple prototype problem is obtained if one fixes the quantity of required output of the product or products made by the enterprise, while prices for the primary products are given. Then the term "use" of resources stands for a choice of a technical process or a combination of processes that meets that requirement within the given constraints. "Best" use is a, or if unique, that choice that meets the requirement at minimum cost of primary inputs.

Economists have differed as to whether this problem belongs in economics. In the twenties the British economist A. C. Pigou stated

> . . . it is not the business of economists to teach woolen manufacturers how to make and sell wool, or brewers how to make and sell beer . . .

This was not the attitude of economists in several other European countries. In particular, there was in the thirties a lively discussion among Scandinavian and German economists concerning models of production possibilities and their use in achieving efficiency within the enterprise. The Nordisk Tidsskrift for Teknisk Ökonomi provided an important medium for these discussions.

Significant contributions[1] were made by Carlson, Frisch, Gloerfelt Tarp, Schmidt, Schneider, Stackelberg, and Zeuthen.

Thus the situation at the end of the thirties was one in which important practical problems in the best use of resources within the enterprise had been neglected by economists in several countries, and had been taken up by only a handful of economists in a few other countries. In addition, the problems were of a kind in which special knowledge possessed by other professions, mathematicians, engineers, managers, was pertinent. One could therefore have expected important new contributions to come from these neighboring professions.

This is precisely what happened, and several times over. Chronologically first was the publication by the mathematician Leonid V. Kantorovich (1939, in Russia) of a 68-page booklet entitled, in translation (1960), "Mathematical Methods of Organizing and Planning of Production." The importance of this publication is due to the simultaneous presence of several ideas or elements, some of which had also been present in earlier writings in different parts of economics or mathematics. I enumerate the elements.

(1) *A model of production* in terms of a finite number of distinct production processes, each characterized by constant ratios between the inputs and outputs specific to the process.

This element has a long history in economics. It is found in Walras (1874, Leçon 41; 1954, Lesson 20), Cassel (1919, Ch. IV), the mathematician von Neumann (1936), Leontief (1936, 1941), all dealing with models of the productive system as a whole. However, the feature most important for our purpose was present only in the classical writers in the theory of international trade[2] and in the models of von Neumann and of Kantorovich. This feature is that the output of one-and-the-same required commodity can in general be achieved by more than one process. The same specified vector of outputs of all required commodities can therefore in general be obtained as the outcome of many different combinations of processes. Two such combinations may differ in the list of processes included - and in the levels of activity assigned to the processes they both use. It is due to this element of choice between alternative ways of achieving the same end result that a genuine optimization problem arises. It is true that Walras also optimized (1954, Lesson 36) on the choice of processes, but from an infinite collection defined by a differentiable production function. It is precisely this choice of a more general collection of processes that delayed the recognition by economists of the applications that are our present topic:

(2) The perception of a wide range of *practical applications* of the model to industries that themselves are sources for the data required by these applications.

[1] For references to these authors and to the Pigou quotation, see Koopmans (1957), p. 185.

[2] See the references to Torrens (1815), Ricardo (1817), Mill (1852), Graham (1923), and others in the survey article by Chipman (1965).

These included the transportation problem to be discussed below, an agricultural problem, and various industrial applications. The definition and collection of available data of a different, more aggregative, kind was also an important element in Leontief's input-output analysis.

(3) The demonstration that with an optimal solution of the given problem, whether of cost minimization or output maximization, one can associate what in Western literature has been called *shadow prices,* one for each resource, intermediate commodity or end-product.

Kantorovich's term in 1939 was "resolving multipliers", which he changed to "objectively determined valuations"[3] in his book of 1959. In general, these valuations are equal to the first derivatives, of the negative of the cost minimum, with respect to the specified availabilities of the goods in question. In mathematical terminology these valuations have also been called "dual variables", in contrast with the activity levels assigned to the processes, which are then called "primal variables". Analogous dual variables occur also in von Neumann's model of proportional growth, with an interpretation as prices in competitive markets.

(4) The identification of a ***separation theorem*** for convex sets due to Minkowski as a mathematical basis for the existence of the dual variables[4].

(5) The ***computation*** of optimal values of the primal and associated dual variables for illustrative examples, and some indications toward calculating such solutions in more complex cases.

Finally, brief but precise explanations of

(6) The interpretation of the dual variables as defining equivalence ratios ***(rates of substitution)*** between different primary inputs and/or different required outputs, and

(7) the additional interpretation of the dual variables as ***guides for*** the coordination of ***allocative decisions*** made in different departments or organizations.

I shall return to (7) below.

Kantorovich's work of 1939 did not become known in the West until the late fifties or early sixties. Meanwhile the transportation model was redeveloped in the West without knowledge of the work on this topic by Kantorovich (1942, reprinted 1958) and Kantorovich and Gavurin (1940, 1949). The Western contributions were made by Hitchcock (1941), Koopmans (memo dated 1942, published 1970; articles of 1949 and 1951 (with Reiter), Dantzig (Ch. XXIII in Koopmans, ed., 1951).

The general linear model was rediscovered and developed by George B. Dantzig and others associated with him, under the initial stimulus of the scheduling problems of the United States Air Force. The term "linear programming" came into use for the mathematical analysis and computational procedures associated with this model. A compact early publication of this

[3] Ob'ektivno obuslovlennye otsenki.

[4] For this purpose von Neumann had used the heavier tool of a topological fixed-point theorem. The dispensibility of this for his purpose was shown later by Gale (1956) and by Koopmans and Bausch (1959, Topic 5).

work can be found in a volume entitled "Activity Analysis of Production and Allocation", edited by Koopmans (195 1). Substantial further developments appeared in such Journals as Econometrica, Management Science, Operations Research, and were brought together in Dantzig's "Linear Programming and Extensions" (1963), a book that was many years in the making. These developments, in which many mathematicians and economists took part, went substantially beyond the earlier work of Kantorovich, in several directions. I note only a few of the extensions to the elements listed above.

(2') *Extension of the range of applications* to animal feeding problems, inventory and warehousing problems, oil refinery operations, electric power investments[5] and many other problems.

(3', 4') Further clarification of the ***mathematical relations between primal and dual variables*** and their extension to ***nonlinear programming*** by[6] Tucker, Gale, Kuhn and others.

This work also traced additional mathematical origins or precursors for the duality theory of linear programming in the work on game theory by von Neumann (1928 and, with Morgenstern, 1944) and by Ville (1938), and in work on linear inequalities by Gordan (1873), Farkas (1902), Stiemke (1915), Motzkin (1936) and others.[7]

(5') The development by Dantzig of the ***simplex method*** for maximizing a linear function under linear constraints (including inequalities) and the further improvements to this method by Dantzig and others.

The simplex method has become the principal starting point for a family of algorithms dealing with linear and convex nonlinear allocation problems. These methods can be set up so as to compute optimal values of both primal and dual variables.

Most important to economic theory as well as application was a further extension of (7) into

(7') analysis of the *role or use of prices* toward best allocation of resources, either through the operation of competitive markets, or as an instrument of national planning.

These ideas, again, have a long history in economics. In regard to competitive markets, they go back at least as far as Adam Smith (1776), and were eloquently restated and developed by Hayek (1945). Important writers on the use of prices in socialist planning were Barone (1908), Lange (1936), and Lerner (1937, 1938). The new element in the work by Koopmans (1949, 1951) and Samuelson (1949, 1966) was the use of the linear model and, in my own case, the attempt to develop what may be called a *pre-*institutional theory of allocation of resources. It was already foreshadowed in the work of Lange and Lerner that hypothetical perfect competition and hypothetical perfect planning both imply efficient allocation of resources - although neither occurs in reality.

* See, for instance, Massé and Gibrat (1957).

[6] See Gale, Kuhn and Tucker (1951), Kuhn and Tucker (1950), and, for a summary, Tucker (1957).

[7] For references see Dantzig (1963), Ch. 2-3.

It therefore seemed useful to turn the problem around, and just postulate allocative efficiency as a model for abstract, pre-institutional study. Thereafter, one can go on to explore alternative institutional arrangements for approximating that model.

I believe that the linear model offers a good foothold for this purpose. First, it makes a rigorous discussion easier. Secondly, the most challenging ***non-*** linearity - that connected with increasing returns to scale - in fact undermines competition. It also greatly escalates the mathematical and computational requirements for good planning. The linear model, therefore, makes a natural first chapter in the theory of best allocation of resources. In its simplest form it leads to the following symmetric relationships between activity levels of the processes and the (shadow) prices of the resources and goods produced:

(7”) (a) Every process in use makes a zero profit,

(b) No process in the technology makes a positive profit,

(c) Every good used below the limit of its availability has a zero price,

(d) No good has a negative price.

These same relationships are a recurrent theme in the first two chapters of Kantorovich’s (1959) book, which also was many years in the making prior to publication. It was subsequently translated into French and English, the latter under the title “The Best Use of Economic Resources.” The gist of the book’s recommendations is that socialist planning can achieve best attainment of the goal set by the planning body through calculations that ensure the fulfillment of these or similar conditions for optimality.

Kantorovich did not go much beyond his earlier remarks on the questions concerning possible use of a price system for decentralization of decisions. This became a major theme, however, in the abstract work of Koopmans (1951, Sec. 5.12), in the work on two-level planning by Kornai and Lipták (1962, 1963, 1965) in relation to planning in Hungary, and in that by Malinvaud (1967) stimulated by experiences with planning in France. The principal computational counterpart of this work was developed by Dantzig and Wolfe (1960, 1961) under the name “the decomposition principle.”

The third chapter of Kantorovich’s book deals with the problem of investment planning to enlarge the production base. The principal emphasis is on the concept of the ***normal effectiveness of capital investment.*** This is a discount rate applied to future returns and to contemplated investments and other future costs, in the evaluation and selection of investment projects. This idea had been proposed earlier by Novozhilov (1939). The point emphasized by Kantorovich is that the prices to be used in calculating returns and costs should be the objectively determined valuations determined by his methods, for the selection to have an optimal result. These proposals were at the time new to the practice of Soviet economic planning. I believe that the principle of the normal effectiveness of capital investment has gained increasing acceptance in Soviet theory and practice since that time. There is an obvious formal analogy with the profitability criterion for investment planning used by the firm in a market economy, using anticipated market prices and the appropriate market rate of interest. The institutional framework contemplated is, of course, fundamentally different. I believe that the underlying pre-institutional optimizing theory is the same.

Summing up, I see two principal merits in the developments I have reviewed so far. One is their initially pre-institutional character. Technology and human needs are universal. To start with just these elements has facilitated and intensified professional contacts and interactions between economists from market and socialist countries. The other merit is the combination and merging of economic theory, mathematical modeling, data collection, and computational methods and algorithms made possible by the modern computer. A genuine amalgam of different professional contributions!

The linear model, followed by the convex nonlinear model, have provided the proving ground for these developments - and - their most conspicuous limitation: The nonconvex nonlinearities associated with increasing returns to scale - i.e., with the greater productivity of large-scale production in many industries - require quite different methods of analysis, and also raise different problems of institutional frameworks conducive to best allocation.

I now proceed to a rather different class of applications of the idea of best allocation of scarce resources. This field is usually referred to as the ***theory of optimal economic growth.*** In most studies of this kind made in the countries with market economies there is not an identifiable client to whom the findings are submitted as policy recommendations. Nor is there an obvious choice of objective function, such as cost minimization or profit maximization in the studies addressed to individual enterprises. The field has more of a speculative character. The models studied usually contain only a few highly aggregated variables. One considers alternative objective functions that incorporate or emphasize various strands of ethical, political, or social thought. These objectives are then tried out to see what future paths of the economy they imply under equally simplified assumptions of technology or resource availability. The principal customers aimed for are other economists or members of other professions, who are somewhat closer to the making of policy recommendations. These may be those engaged in making more disaggregated optimizing models of growth that incorporate numerical estimates of technological or behavioral parameters. (I shall return to this field of "development programming" below.) Or the hoped-for customers may be policy economists who may find it useful to have the more abstract ideas of this field in the back of their mind when coping with the day-to-day pressures for outcomes rather than criteria.

The question of the clientèle is even more baffling when the problem concerns growth paths for time spans covering several generations. What can at best be recommended in that case is the signal the present generation gives, the tradition it seeks to strengthen or establish, for succeeding generations to take off from.

The classic in the optimal growth field is a paper published in 1928 by Frank Ramsey, known also as the author of equally fundamental papers on the foundations of mathematics and on subjective probability. His definition of "best" involves the maximization of a sum (or integral) of utility flows to be derived from future consumption. Using a continuous time variable, Ramsey's choice of objective function is a limiting case of a broader class of functions which I shall consider first,

$$U = \int_0^\infty e^{-\varrho t} u(c_t)\,dt,\ 0 < \rho, \quad \left\{ \begin{array}{l} \text{the objective function} \\ \text{representing generations.} \end{array} \right.$$

Here c_t denotes the aggregate consumption flow as of time t, and $u(c)$ is a utility flow serving as an evaluating score for the consumption flow c. One chooses the function $u(c)$ so as to increase with c but at a decreasing rate $\frac{du}{dc}$ as c increases. This expresses that at all times "more is better", but less so if much is already being enjoyed (see Figure 1). The effect on the allocation of consumption goods between generations is similar to the effect of a progressive income tax on spendable incomes among contemporaries.

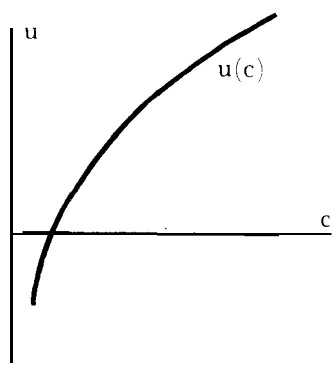

Fig. 1

We shall call the exponentially decaying factor $e^{-\varrho t}$, $\varrho > 0$, ***the discount factor*** for utility. It diminishes the weight given to future utility flows in the summation of the entire utility flow over all the future to form a total score U. The weight is smaller the larger the ***discount rate*** for ***utility,*** ϱ, and the further one looks into the future. On ethical grounds Ramsey would have none of this. I shall take the view that the important question of discounting utility - or for that matter any other aspect of the choice of the objective function - should not be settled entirely on a priori grounds. Most decision makers will first want to know what a given objective function will make them do in given circumstances. I shall therefore hold $\varrho > 0$ for this first exploration,[8] and turn to the mathematical modeling of the "circumstances" in terms of technological and resource constraints on the consumption and capital variables.

One "resource" is the labor force. It need not enter the formulae because it will be assumed to remain inexorably constant over time. The only other resource is an initial capital stock denoted k_0, historically given as of time $t = 0$. The "use" at any time t of labor and of the then capital stock k_t consists

[8] For an objective function implying a variable discount rate that depends on the path contemplated see Koopmans (1960), Koopmans. Diamond and Williamson (1964) and Beals and Koopmans (1969).

of two steps. The first and obvious step is to achieve at. all times the highest net output flow $f(k_t)$ that can be produced by the labor force, using the capital stock fully and to best effect. The form given to the function $f(k)$ summarizes and simplifies broad technological experience. It specifies $f(0) = 0$ ("without capital no output"), $f(k)$ initially increasing with k but at a diminishing rate $\frac{df}{dk}$, in such a way that from some point $\hat{k}$ of capital saturation on, $f(k)$ decreases because depreciation rises more steeply than gross output. (See Figure 2.)

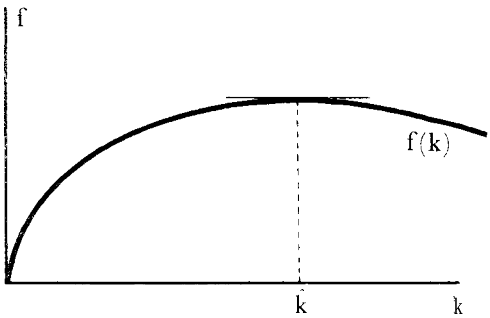

Fig. 2

In all this the product flow $f(k)$ is regarded as consisting of a single good, which can be used as desired for consumption or for adding to the capital stock,

$$f(k_t) = c_t + \frac{dk_t}{dt}, \quad \text{the allocation constraint.}$$

To determine this allocation for all t is the second step. This is done "best" at all future points of time if the total score U is thereby maximized.

It might seem as if this constrained maximization problem is quite different in mathematical structure from those discussed before. This is not the case. The main difference is that the discussion has shifted from a vector space to a function space, using conventional notations not designed to reveal the common structure of the two problems. In particular, as long as the crucial convexity assumptions are maintained, interpretations in terms of shadow prices remain valid.

The problem for $\varrho > 0$ was solved independently by Cass (1966), Koopmans (1965, 1967), Malinvaud (1965, 1967), thirty-five years after Ramsey. Without proof I indicate the nature of the solution in Figure 3. In the diagram on the left, the ***abscissa k*** is set out along the ***vertical*** half-axis the ***odinate*** $y = f(k)$ along the ***horizontal*** half-axis pointing left. For given $\rho > 0$, find the unique point $\hat{k}(\rho)$ on the curve $y = f(k)$ in which the slope $\frac{df}{dk}$ equals ρ. Then, if the initial capital stock k_0 should happen to equal $\hat{k}(\rho)$, the optimal capital path remains constant, $k_t = \hat{k}(\rho)$, over all the future. For any initial stock k_0 less than or larger than $\hat{k}(\rho)$, the optimal path shows a monotonic and asymptotic approach to $\hat{k}(\rho)$. All this is illustrated in solid lines in the top right diagram in Figure 3. The lower right diagram shows the corresponding optimal consumption path c_t, which approaches the asymptotic level $c(\rho) = f(k(\rho))$.

First, a more realistic model should incorporate estimates of the relations describing the response of reproductive behavior to levels income, education, housing, medical care, and other causative variables. In circumstances where the resulting path of population does not seriously reduce per capita income below what could be achieved by a more direct population policy, no further action would be required. Where this is not the case, the processes whereby reproductive behavior can be influenced, and in particular the relation between resource inputs into these processes and the responses to them (prompt and delayed) need to be incorporated in the model. Considerations of this kind have been introduced into optimal population models by Pitchford (1974).

The second refinement concerns the optimality criterion. Situations occur in which what can be achieved by population policies can only diminish but not prevent a lowering of the per capita consumption sustained by domestic resources, for an extended period ahead. In such cases a realistic view will recognize a degree of exogeneity in the future path of population for some time to come. One may then want to explore optimality criteria that extend to those as yet unborn children whose birth had better, but cannot, be prevented the same consideration as to those already born.

We have considered two broad fields of application for optimization models. One comprises the detailed and data-oriented optimization of the decisions of the enterprise or public agency- and also the coordination of such decisions through a price system, through centralized planning and management, or both. The other is the more speculative study of alternative aggregate future growth paths for an entire economy.

In conclusion I want to make some remarks about the growing field of ***development programming,*** in which the two strands of thought are being combined and merged. One early step in this development was the construction of a mathematical programming model for an economy as of some future year, including investments and the flow of aid in the intervening period as decision variables. An example is the study of the economy of Southern Italy by Chenery, writing with Kretschmer (1956) and with Uzawa (1958). An evaluative description of experiences with Hungarian economic planning along these lines was written by Kornai (1967). Later studies, such as that of the Mexican economy by Goreux, Manne and coauthors (1973), envisaged a sequence of future years. In most of these studies data availabilities determined the use of Leontief's input-output framework for representing the production possibilities of the economy as a whole. Policy choices and optimization were introduced where data so permitted. One example is the choice between domestic production versus imports paid for by exports in the Southern Italy study. Others are the sectoral detail in the Hungarian studies, and concentration on the energy and especially electric energy sectors in the Mexican one.

A weakness in the treatment of consumption in optimal growth models, noted by Chakravarty (1969), is the lack of continuity between consumption levels in the past and those recommended by otherwise reasonable looking optimality criteria for the near future. One remedy proposed by Manne