MACROECONOMICS AND LINEAR PROGRAMMING

By Ragnar Frisch

> The biggest problems will probably be encountered in the social and economic field and these problems must be studied in the same thorough-going way as those pertaining to technology and the natural sciences.
>
> Norwegian Prime Minister *Einar Gerhardsen* in an address February 1956 to a conference on »Technical development and the future».

1. How is linear programming related to macroeconomics?

One of the most fascinating aspects of economics is the great variety of mental activities that are involved in its study. In one end of the spectrum we find painstaking fact digging, in the middle we find broad philosophical thinking and in the other end we find heavy mathematical reasoning which becomes indispensable when we shall keep track of a multitude of logical elements simultaneously.

The evolution of the study of one large and important field of economics may be taken as an illustration in point. A generation or two ago it was customary to think of monetary theory as a more or less watertight compartment that was separated from general economic theory (that revolved around the analysis of relative prices taken in a very broad sense). And still more sacredly it was kept apart from the principles and methods of public finance. It was thought to be too dangerous to let the minister of finance become conscious of the enormous possibilities which would be opened up for him if he extended his activities into the monetary field and began to consider the more *comprehensive* problem: that of steering the whole economy, or at least a very large part of it, simultaneously.

Foremost in the group of economists who inaugurated a new approach in these matters, is Professor Erik Lindahl, who in this respect (and others as well) kept up the tradition from Knut Wicksell. In the famous works »Penningpolitikens mål» from 1924 and »Penningpolitikens medel» from 1929 Lindahl embarked upon a systematic study of how monetary theory can be geared into general economic theory and both tools can be used to discuss the steering of the economics in *a balanced way.*

Subsequently the evolution towards a global theory and a global formulation of the targets of economic policy has proceeded by leaps and bounds. The latest phase of this development is the application of the mathematical tool known as linear programming to economic policy. The present paper will be concerned with this development. The formulation of the problem of economic policy in terms of linear programming is simple enough. Indeed, it involves no more than elementary college algebra. But the solution is not so simple. As I shall be concerned also with methods of solution, a rather technical form of the last part of the paper was inevitable.

The mathematical essence of linear programming is that we consider a number of variables that are subject to the condition of being *non-negative,* and furthermore such that certain linear *functions* of them shall be non-negative, and finally such that certain linear *equations* between the variables shall be fulfilled. The set of values of the variables that satisfy these conditions is called *the admissible region.* A more detailed discussion of the concept is given in Section 2.

A multitude of examples from the macroeconomic field immediately come to mind. For instance if the variables are quantities produced in a number of production sectors, or quantities consumed in a number of specified consumer groups, or quantities used for investment in fixed capital, or quantities imported for use in the production sectors or by the consumer groups, or quantities of labour engaged in the various production sectors etc., etc., these magnitudes will in a concrete problem as a rule be subject to the condition of being *non-negative.* And under simplifying assumptions they will be related by a great number of *linear equations,* for instance equations derived from an interindustrial input-output analysis, or from a study of Engel-elasticities in the consumer groups, or from a study of how imports and labour depend on the level of activities in the various production sectors.

But this is not all. There will often be a *linear function* of the variables or of some of the variables that must by the nature of the concrete problem be non-negative. There may even be several such linear functions that must

be non-negative. For instance, if it is required that there shall be produced given quantities for consumption or for investment, it will, by the classical methods of input-output analysis, be possible to figure out how high *the total level of production* in each sector must be when all direct and indirect effects are included. Each such level of production will — when constant production coefficients are assumed — be a *linear* function of the prescribed consumption and investment elements; and the coefficients of these linear functions can be computed. However, in practice under given short run conditions there will be a definite upper bound to the level of production in each sector, due to the *fixed capital equipment* which exists in the sector. Hence for each production sector where capital capacity is an important element, there exists a certain linear function of the consumption and investment elements which must be less than or equal to the existing capital capacity in the sector. In other words, there exists in each such sector a well defined linear function of the investment and consumption elements that must be non negative.

In a similar way we may derive bounds by considering the available *labour force* and the nature of the immobility of labour.

Similarly we may derive bounds by considering the highest permissible strain on the *foreign exchange balance.*

Or we may consider nearly any sort of bounds that expresses a humanitarian, social or political desideratum.

The joint effect of all such bounds define the admissible region.

When the admissible region is determined, the formulation of the goals for economic policy is completed by defining a *preference* function that we want to maximize, subject to the condition that the variables shall stay within the admissible region. The preference function may for instance be formulated simply as the gross national product. That is to say, we may formulate as a target for economic policy to maximize gross national product, subject to the conditions expressed by the admissible region. Or the preference function may express a compromise between making the gross national product high and reducing the strain on the foreign exchange balance. Or it may express a manysided compromise where also considerations on the desirability of certain types of long run investment comes in. And so on. In any case will the preference function be formulated as a linear (or some times a non linear) function of some of or all the variables, the coefficients in the function being numerically well defined.

The structure of the admissible region and the shape of the preference function will, of course, be very different in the case of a short run and in

that of a long run development program, but I shall not insist on this here.

How to determine the coefficients in the preference function is a question by itself. The decisive influence on these coefficients must come from the responsible politicians or political bodies, not from the programming technician. But, on the other hand, it is only the programming technician who will understand the meaning of the coefficients sufficiently well to be able to reckon with them if the preference function contains many terms. An indirect method must therefore be used. This problem is discussed in Section 3.

2. *Formulation of the mathematical problem*

In standard form the linear programming problem can be formulated mathematically as follows.

Consider $(n + m)$ real variables $x_1, x_2 \ldots x_{n+m}$ satisfying m linearly independent linear equations. The equations may be written in the standard form

$$(2.1) \qquad a_{io} + \sum_{j=1}^{n+m} a_{ij}x_j = 0 \qquad (i = 1,2 \ldots m)$$

where the matrix a_{ij} is of rank m (and none of the variables is lacking in all the equations). The number of degrees of freedom is consequently n and it is always possible at least in one way to express all the variables in terms of a set of n basis variables, linearly independent amongst themselves. Let

$$(2.2) \qquad x_u, x_v \ldots x_w$$

be such a basis set. The equations may then be written in the basis form

$$(2.3) \qquad x_j = b_{jo} + \sum_{k=u,v\ldots w} b_{jk}x_k \qquad (j = 1,2 \ldots n + m)$$

where the b_{jo} and b_{jk} are constants. Obviously

$$(2.4) \qquad b_{jo} = 0 \text{ and } b_{jk} = \left\{\begin{matrix} 1 \text{ if } k = j \\ 0 \text{ otherwise} \end{matrix}\right\} \text{ when } j = u, v \ldots w$$

If the equations (2.3) are taken for all $j=1, 2 \ldots n + m$, we get a system of equations that are lineraly dependent, but if we take (2.3) only for[1] $j=1, 2 \ldots)u, v \ldots w(\ldots n + m$, we get a system of equations that are linearly independent. More precisely: If the coefficients b_{jo} and b_{jk} have any values whatsoever, m equations of the form (2.3) for $j=1, 2 \ldots)u, v \ldots w(\ldots n + m$ are always linearly independent.

[1] We use the inverted parenthesis) . . . (to denote "exclusion of".

We consider a linear *preference function*

$$f = \pi_o + \sum_{j=1}^{n+m} \pi_j x_j \tag{2.5}$$

where the π are given constants positive, negative or zero. If we introduce the expressions for the dependent variables from (2.3), the preference function assumes the form

$$f = p_o + p_u x_v + p_v x_v + \ldots + p_w x_w \tag{2.6}$$

where the p_k ($k=0, u, v \ldots w$) are known constants positive, negative or zero. It does not restrict generality if we confine our attention to the form (2.6).

The linear programming problem is the problem of determining that one or those sets of values of the variables that will maximize (2.6) subject to the following two sets of conditions: In the first place the *equations* (2.3), and in the second place the non-negativity conditions expressed by the *inequalities*

$$x_j \geqq 0 \qquad (j = 1,2 \ldots n+m) \tag{2.7}$$

A great variety of problems can be reduced to this form. If there should be one of the variables that is *not* subject to the non-negativity condition, this variable could be eliminated and the problem reduced to one with the same number of degrees of freedom but one equation and one variable less. A problem containing any linear inequality, that is, the condition that a given linear function of the variables shall be non-negative, can be reduced to the above form simply by taking the value of the linear function in question as a new magnitude that is entered in the list of variables.

The inequalities (2.7) introduce *discontinuities* in the boundary conditions, and this makes the method of Lagrange multipliers, which serves so well in many other maximum problems with side conditions, inapplicable here.

The now classical method for handling the linear programming problem is the simplex method due to George B. Dantzig. At the Oslo University Institute of Economics considerable effort has been made to handle the problem in a different way in the hope of finding one or some methods that may be more advantageous in cases with a great number of variables, particularly in problems of the type occurring in macroeconomic planning. In a national planning problem of some size one may easily run into several hundred variables and perhaps a hundred or more degrees of freedom. And with modern computing outfit it is not unrealistic to be prepared for a

higher generation of problems with tens of thousands of variables and several thousand degrees of freedom.

It is an easy form of mental exercise to imagine the solution in the form of iteration processes of one form or another. They may look nice on paper, but in most cases such methods do not converge, or if they do converge in principle, the convergency will in general be so slow as to make the method entirely useless in practice. At the Oslo Institute we have accumulated a considerable junk-pile of such methods. Some of the methods we have used, do, however, contain features that may make them useful under certain circumstances. Some aspects of methods which have actually been used successfully on small or medium sized examples, are described in a number of mimeographed memoranda from the Oslo Institute of which the following might be mentioned here:

[1]	21 June 1954	Methods of solving linear programming problems. Synopsis of a lecture to be given at the International Seminar on Input-Output Analysis, Varenna (Lake Como) June—July 1954.
[2]	18 October 1954	Principles of linear programming. With particular reference to the double gradient form of the logarithmic potential method.
[3]	29 March 1955	A labour saving method of performing freedom truncations in linear programming. Part I.
[4]	13 May 1955	The logarithmic potential method of convex programming. With particular application to the dynamics of planning for national development. Synopsis of a communication to be presented at the international colloquium of econometrics in Paris 23—28 May 1955.
[5]	17 October 1955	The multiplex method for linear programming.
[6]	3 January 1956	The logarithmic potential method for linear programming formulated with a view to electronic computation.
[7]	10 January 1956	Macroeconomics and linear programming.

It should always be remembered that any mathematical method and particularly methods in linear programming must be judged with reference

to the type of computing machinery available. In all our work we have been guided by the possibilities and desiderata in a situation where the available equipment consists of desk machines or IBM 602 A (or its electronic improvement 626, or similar types of calculating punches) or electronic automatic computors with a small high speed memory, as for instance the Oslo machine NUSSE (whose name academically stands for »Norwegian Universal etc.», but is better understood when referred to the meaning of the Norwegian word nusse which can appropriately be translated as »small cute girl»). Our outlook may perhaps be changed when we get used to the super modern, high capacity electronic computor that will be available here from the middle of next year.

The present paper describes — with a minimum of proofs — an approach which we have found promising.

As a background for all methods of solving linear programming problems must be kept in mind the general properties of the solutions of such problems. In short they may be summarized as follows.

The set of points that are optimal — i. e. that belong to the admissible region and are such that in this region it is impossible to find any point capable of producing a value of the preference function that is higher than the value produced in the point considered — form a coherent linear manifold belonging to the *boundary* of the admissible region, i. e. belonging to the point set of the admissible region where at least one of the variables is exactly zero. The dimensionality of this pointset, that is the number of degrees of freedom in it, may be any of the numbers $\delta = 0,1 \ldots n-1$, where n is the number of degrees of freedom in the formulation of the problem. The case $\delta = 0$ means that there is only one well defined *corner*[1] on the boundary of the admissible region where the preference function assumes its maximum value, the case $\delta = 1$ means that the maximum is reached all along an *edge* that connects two corners, etc.

Whatever the dimensionality of the optimum pointset there exists at least *one* corner with optimal properties, i. e. at least one optimal point which is such that in this point at least n of the variables are zero, these n variables being such that their vanishing is exactly sufficient to determine the point. If $\delta = 0$, there exists one and only one optimal corner, if $\delta = 1$ there exist exactly two optimal corners, and in general if δ is any given number $(< n)$, there exist $\delta + 1$ optimal corners.

[1] I. e. a vertex.

3. *Practical fixation of the bounds and the preference function*

How to fix the bounds and how to determine the coefficients of the preference function are important practical problems. They are indeed so far-reaching that they lead us into a general consideration on the line of demarcation between the work of the politician and that of the scientist. Expressed briefly and therefore necessarily without complete precision, we can say that the politician must introduce the human evaluations, the social value judgements, while the task of the scientist is objectively to find out what the factual situation is and what the inherent tendencies for change are and what consequences *could be expected* if one decided to put into effect such and such measures. In this work the scientist will simply have to take as data the goals themselves and the social value judgements back of them.

If we scrutinize this distinction closer, we will, of course, — as always when it is a question of distinctions of principle — see that there can be marginal cases which are difficult to decide. In the last resort we will perhaps have to retain only this formulation: The goals and the social value judgement are what the scientists *do not desire to take up for analysis.* It is that part of the problem which is too difficult or too vague to be amendable to exact scientific methods. Therefore, to a certain extent the distinction becomes relative and may change as we change the purpose of the analysis or we get at disposal new tools of analysis or new factual information. For all practical purposes the distinction between the task of the politicians and that of the scientists is, however, clear enough.

The analytical system back of a planning work will as a matter of fact follow very closely common sense. It is nothing but common sense put into system. When it does not appear in this way to many who consider themselves »practical men», it is only because the scientist in order to save time and effort must use a terminology that does not belong to every day language. And also because he has to use an analytical apparatus of a certain size.

Common sense tells us that if one wants to steer the evolution of a country, one must first consider what the situation actually is and next decide what one wants it to be, and then look around and see what possibilities exist of bringing the situation from what it is to what one wants it to be. These are the main lines of the theoretical set up.

In all underdeveloped countries — and what country could not be called underdeveloped in one sense or another — the concept of *time* is particularly important. In how long spans of time shall we reason when we are doing

national economic planning? Recently I had an opportunity of working on the methodology of economic planning in India, and I well remember how India's Prime Minister Mr. Nehru in a discussion emphasized the enormous difference between the problem as it appears in the United States and in India: In the United States the problem was how one could bring the most refined technical gadgets, let us say in the construction of refrigerators, into general use. In the Indian population the urgent problem was how to keep *hunger* away from the population and how the country could work up a reserve stock of food grains and other goods fundamental for feeding the population, so that one did not risk to be forced into a situation where food grains *had to* be imported at any cost. When one remembers the situation in India some years ago, this distinction between the two types of problems is obvious and indeed very realistic. But when it appears that we can in India now safely disregard the problem of technical refinement in refrigerators for the population, it is only because of the *time horizon* one chooses to adopt. When the responsible Indian politicians — rightly — look at the question as they do, it is because they reason within a horizon which is wide enough to make it possible to solve the nutritional problem safely within this span of time, but not wide enough to make it possible to introduce technical refinements in refrigeration for the bulk of the Indian population. This refrigerator problem too *will* present itself in India but only at such a distant future time that we do not need to bother about it now.

How much weight shall be put on the first, the most burning problems and how much on the long views extending into the future, is a matter of judgement. In order to form an opinion on this, the responsible politicians must so to speak already have *guessed* what the solution would be of an imaginary analysis of an enormous problem where all possible details in the present and all possibilities of the future were specified. So we have here an example where one has brushed aside the distinction between the social value forming politician and the objectively working scientist. The politician must, whether he wants it or not, act in both capacities. It is, however, only in the very first phase of the analysis where such a compromise has to be made. For the further study of the complex of problems that has been circumscribed by »the horizon of imaginations of the politician» we can fully and with the greatest efficiency apply the principle of division of labour between the politician and the scientist.

This must be the background of the determination of the coefficients expressing the bounds and the preference function. Through interviews between the responsible politicians and the programming technicians the

latter must try to settle on a set of coefficients that will come the closest possible to giving a correct picture of the value judgements of the responsible politicians.

At the Oslo Institute we have experimented to some degree with the type of interviews necessary to formulate the coefficients in question. First consider the preference function.

Its coefficients must be fixed in an *indirect* way by putting up alternatives of choice and noting the *ranking order* which the politicians or groups of them will prefer. Different schemes are conceivable for obtaining such a ranking order and translating it into preference coefficients. Some of them are discussed in [7].

Whatever values emerge for the coefficients in the preference function, this function can by itself never produce an insolvable programming problem. In this respect the *bounds* are different. The bounds are the results of the impositions of linear inequalities derived, say, from technical considerations on the capacities of the fixed capital outfit in the various sectors, or the limitations in the labour force or its immobility, or derived by the need for economizing the use of foreign exchange, or the need for holding the production of or import of certain essential consumption goods above certain physical levels of existence etc. There is no end to the kind of bounds that may be imposed for technical or humanitarian or political reasons as mentioned in Section 1.

It will always be well to exert the *largest possible moderation* in imposing bounds, i. e. one should reckon with production capacities pressed to the highest conceivable levels, and introduce boundary levels of consumption as low as is by any means permissible or preferably no such bounds at all, etc. If we do not show such moderation, if on the contrary all sorts of exigencies are allowed full play, we will almost certainly run into a contradiction, i. e. get a situation where *no admissible region exists,* and consequently the programming problem has no solution.

If any particular element in the program appears as extremely important, for instance the assurance of a basic standard of consumption for a certain essential consumption good, it is adviceable to formulate the corresponding *bound* in the weakest permissible way and instead let this consumption element enter into the *preference function* with an extremely high weight.[1]

[1] If this is done, we may run into the following difficulties: It may be appropriate to attach an extremely high weight to the consumption element as long as the consumption is low, but it may not be appropriate to do so when this consumption is high. This difficulty can be handled by using a non-linear preference function. As long as the bounds are linear, the admissible region is convex and this is mathematically the essential point. We shall, however, not enter further into this question here.

By so doing the element in question will be sure to influence the final solution strongly and *if at all possible,* it will be brought up to a level above the minimum imposed.

In spite of all moderation used in the formulation of bounds, it may turn out that no admissible region exists. In this case we must look for a *reformulation* of the problems and it becomes necessary to consider in a more systematic way what sorts of reformulations are possible with a minimum of computational costs. This is discussed in [7].

4. Solving linear equations and inverting matrices when high speed memory is limited

An essential point in any method of solving linear equations and inverting matrices is the number of multiplications and/or divisions which it involves. If the work is done by some method more or less akin to the Gaussian algorithm, the number of multiplications and/or divisions involved in a one way solution — that is with one given set of numbers in the right member — is of the order $\frac{n^3}{3}$ where n is the number of unknowns. A complete inversion of the matrix involves a work of the order n^3. It is a remarkable fact that by this method the work involved in a complete inversion is only 3 times as large as that involved in a one-way solution, the factor 3 being independent of n.

It is, however, not only the number of multiplications and/or divisions that count, but also the number of figures that need to be taken down (either in written form or punched on cards or entered, say, on magnetic tape or drum).

And finally we must consider the maximum number of figures that need to be *stored* at any time during the operation. This point is particularly important when the work is done on an automatic computor with a high speed memory of limited capacity. It is possible at some expense of the number of multiplications and/or divisions to shape the elimination process in such a way as to reduce the maximum number of numbers stored. This problem is discussed in [7].

5. Directional techniques

In linear programming work with the Oslo Institute methods we are often confronted with the problem of moving from an initial point in a *direction*

which is determined in some specific way. It is convenient to consider certain general aspects of such directional techniques. In so doing we shall for generality assume that the initial point x_k° (k=u, v . . . w) may be inside or outside the admissible region or on its boundary. That is, any of the variables x_j (j=1, 2 . . . n + m) may be positive, negative or zero.

A direction may be defined by fixing a *directional point* x_k' (k=u, v . . . w) at which we aim from x_k° (k=u, v . . . w). The increments in the basis variables as we move from the initial point to the directional point are

$$d_k = x_k' - x_k^\circ \qquad (k = u, v \ldots w) \tag{5.1}$$

The direction may be defined either through the directional increments d_k or through the values x_k' of the basis variables in the directional point. One of these sets of data is equivalent to the other by (5.1).

For *any* of the variables — basis variables or independent ones — the directional increment is defined by

$$d_j = x_j' - x_j^\circ \qquad (j = 1,2 \ldots n + m) \tag{5.2}$$

If the directional increments for the basis variables are given, those for the other variables can be computed by

$$d_j = \sum_{k=u,v,\ldots w} b_{jk} d_k \qquad (j = 1,2 \ldots n + m) \tag{5.3}$$

This follows from (2.3).

We shall now consider two more general ways of fixing a direction, namely the *completely incremental method* and the *moment method.*

In the completely incremental method we select n variables — basis variables or dependent variables — equal in number to the degrees of freedom and prescribe the directional increments for these selected variables. We assume that the selection is made in such a way that the n x n submatrix

$$b_{ik} \qquad \begin{pmatrix} i = r, s \ldots t \\ k = u, v \ldots w \end{pmatrix} \tag{5.4}$$

is non-singular, where Nos. i=r, s . . . t are the selected variables. With given d_i (i=r, s . . .t) the linear system

$$\sum_{k=u,v\ldots w} b_{ik} d_k = d_i \qquad (i = r, s \ldots t) \tag{5.5}$$

will then have a unique solution in the d_k (k=u, v . . . w). And once the basis direction increments are known, all the other increments are determined by (5.3).

In the moment method we also prescribe the directional increments (or the directional values) for a certain number of selected variables, but now this number may be *less* than the number of degrees of freedom. To make up for the freedom which this will leave, we add the specification that the basis direction vector d_k shall be a linear form in the boundary vectors for

the selected variables. More precisely: Let Nos. i=r, s . . . t be ν ($\leqq$ n) selected variables for which the directional increments d_i (or the directional values x_i') are given. The selected set is assumed such that the boundary vectors

(5.6) $\quad b_{rk}, b_{sk} \ldots b_{tk}$

are linearly independent over the field of variation k=u, v . . . w. We then specify the movement by requiring that the directional vector d_k shall belong to the manifold unfolded by the vectors (5.6) that are orthogonal to the boundary planes, i. e. we shall have

$$(5.7) \qquad d_k = C_r b_{rk} + C_s b_{sk} + \ldots + C_t b_{tk} \qquad (k = u, v \ldots w)$$

where $C_r, C_s \ldots C_t$ is a set of constants independent of k.

If we multiply (5.7) by b_{ik} (i=r, s . . . t) and perform a summation over k, we get

$$(5.8) \qquad M_{ir}C_r + M_{is}C_s + \ldots + M_{it}C_t = d_i \qquad (i = r, s \ldots t)$$

where the moments are defined by

$$(5.9) \qquad M_{\alpha\beta} = \sum_{k = u, v \ldots w} b_{\alpha k} b_{\beta k}$$

(5.8) is a linear system of ν equations to determine the ν unknowns C. The system is non-singular because of our assumption about the linear independency of the vectors (5.6) The system is furthermore symmetric and positive definit, i. e. of the same sort as we encounter in linear regression analysis. When the C are determined through (5.8), the d_k are determined by (5.7) and finally the d_j by (5.3).

When a direction from the initial point x_k° (k = u, v . . . w) is fixed in some way, we may move in that direction a *shorter or longer distance,* not necessarily so far as to get exactly to the directional point we considered originally. During the movement the values of all the variables will change linearly according to the formula

$$(5.10) \qquad x_j = x_j^\circ + \lambda d_j \qquad (j = 1,2 \ldots n + m)$$

where λ is a parameter generating the movement. The value $\lambda = 0$ gives the initial point and $\lambda = 1$ the directional point. If the initial point is in the interior of the admissible region, we may for instance decide to go as far as we can without breaking out of the admissible region. Or the length of the movement may be decided upon by other considerations.

6. *Truncation techniques*

We shall often have to consider not only the process of moving from an initial point to a determinate point satisfying certain specifications as we did in the previous section, but we may be concerned with the more general problem of *determining a new linear manifold* that shall satisfy certain specifications that are less in number than the degrees of freedom in the problem. In particular we may be interested in determining the form of the basis equations in a new *truncated space.* The idea of an initial point is now without avail, we are indeed not concerned with specific points at all or with specific movements but only with the shape of certain linear manifolds.

We begin by selecting a subset of the variables, let it be

(6.1) The μ dependent variables Nos. r, s . . . t and the ν basis variables Nos. R, S . . . T,

where

$$(6.2) \qquad \mu + \nu \leqq n \qquad 0 \leqq \mu \qquad 0 \leqq \nu$$

We assume that the selection has been made in such a way that the rank of the μ x (n—ν) matrix

$$(6.3) \qquad \begin{matrix} b_{ru}\, b_{rv} \ldots) & b_{rR} & b_{rS} \ldots b_{rT} & (\ldots b_{rw} \\ b_{su}\, b_{sv} \ldots) & b_{sR} & b_{sS} \ldots b_{sT} & (\ldots b_{sw} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_{tu}\, b_{tv} \ldots) & b_{tR} & b_{tS} \ldots b_{sT} & (\ldots b_{tw} \end{matrix}$$

is exactly μ, μ being by (6.2) not greater than n—ν. ((6.3) is a stronger condition than a corresponding condition imposed *before* the ν columns of the basis variables Nos. R, S . . . T had been omitted).

Having made such a selection, we want to investigate the linear manifold defined by the $(\mu + \nu)$ conditions

$$(6.4) \qquad \underbrace{x_r = x_s = \ldots = x_t}_{\mu} = 0 \qquad \underbrace{x_R = x_S = \ldots = x_T}_{\nu} = 0$$

The condition (6.4) leave us with n-$(\mu+\nu)$ degrees of freedom. That is, amongst the original basis variables Nos. k=u, v . . . w there exists a subset of n-$(\mu+\nu)$ variables Nos. U, V . . . W such that this set is a linearly independent set and can therefore be used to generate the remaining n-$(\mu+\nu)$ degrees of freedom. As a rule such a set can be chosen in different ways. Since we have assumed (6.3) to be of rank μ, there exists at least one set of n-$(\mu+\nu)$ variables that can be chosen for our purpose, namely the set obtained from

u, v . . .) R, S . . . T(. . . w by leaving out the μ variables Nos. H=A, B . . . C, where A, B . . . C is a set such that the determinant

$$(6.5) \qquad |\, b_{iH} \,| = \begin{vmatrix} b_{rA} & b_{rB} & \dots & b_{rC} \\ b_{sA} & b_{sB} & \dots & b_{sC} \\ \dots & \dots & \dots & \dots \\ b_{tA} & b_{tB} & \dots & b_{tC} \end{vmatrix} \qquad \begin{pmatrix} i = r, s \dots t \\ H = A, B \dots C \end{pmatrix}$$

is different from zero.

If we want to, we can express all the remaining variables, i. e. those that have not been put equal to zero, namely Nos. j=1, 2 . . .) r, s . . . t, R, S . . . T(. . . n + m in terms of the n-($\mu + \nu$) basis variables Nos. K=u, v . . .) R, S . . . T, A, B . . . C(. . . w. For brevity we denote these basis variables

$$(6.6) \qquad K = \underbrace{U, V \dots W}_{n-(\mu+\nu)} = u, v \dots) \underbrace{R, S \dots T}_{\nu} (\dots) \underbrace{A, B \dots C}_{\mu} (\dots w$$

Thus, Nos. U, V . . . W is a subset of the original basis variables Nos. u, v . . . w, and it is a linearly independent subset. In short it is simply a *truncated* set of basis variables.

If b and b' are the coefficients of the basis equations in the original and the new form respectively (before any of the variables have as yet been put equal to zero) we have, denoting the inverse of (6.5) b^{-1}_{Hi} and for shortness p_k by b_{ok}

$$(6.7) \qquad b'_{Hi} = b^{-1}_{Hi} \qquad \begin{pmatrix} H = A, B \dots C \\ i = r, s \dots t \end{pmatrix}$$

$$(6.8) \qquad b'_{HK} = - \sum_{i = r, s \dots t} b^{-1}_{Hi}\, b_{iK} \qquad \begin{pmatrix} H = A, B \dots C \\ K = 0, u, v \dots) AB \dots C (\dots w \end{pmatrix}$$

$$(6.9) \qquad b'_{ji} = \sum_{H = A, B \dots C} b_{jH}\, b^{-1}_{Hi} \qquad \begin{pmatrix} j = 0, 1, 2 \dots) u, v \dots w, r, s \dots t \\ (\dots n + m \\ i = r, s \dots t \end{pmatrix}$$

$$(6.10) \qquad b'_{jK} = b_{jK} + \sum_{H = A, B \dots C} b_{jH}\, b'_{HK} \begin{pmatrix} j = 0, 1, 2 \dots) u, v \dots w, r, s \dots t \\ (\dots n + m \\ K = 0, u, v \dots) A, B \dots C (\dots w \end{pmatrix}$$

$$(6.11) \qquad b'_{jK} = b_{jK} - \sum_{= r, s \dots t} b'_{ji}\, b_{iK} \qquad \begin{pmatrix} j = 0, 1, 2 \dots) u, v \dots w, r, s \dots t \\ (\dots n + m \\ K = 0, u, v \dots) A, B \dots C (\dots w \end{pmatrix}$$

The number of multiplications and/or divisions involved — not counting checks and assuming that the μ order inversion is performed by a method — for instance the Gaussian elimination algorithm — that involves a work of order μ^3, will be as indicated in tab. (6.13).

TABLE (6.13) Number of multiplications and/or divisions involved in a shift of the content of the basis set.

Formula No.	Work involved when the μ dependent variables Nos. r, s . . . t and the ν basis variables Nos. R, S . . . T are put equal to zero	Additional work when no degrees of freedom are taken out, but the previous basis variables Nos. A, B . . . C are taken out of the basis set and instead the μ previously dependent variables Nos. r, s . . . t are taken into the basis set
(6.7)	μ^3	0
(6.8)	$\mu^2 (n - \mu - \nu + 1)$	$\mu^2 \nu$
(6.9)	0	$\mu^2 (m - \mu + 1)$
(6.10)	$\mu (m - \mu + 1) (n - \mu - \nu + 1)$	$\mu\nu (m - \mu + 1)$
Total	$\mu^3 + \mu (m + 1) (n - \mu - \nu + 1)$	$\mu (\mu + \nu) (m + 1) - \mu^3$

When the b'_{jK} (j=1, 2 . . .) u, v . . . w, r, s . . . t(. . . n + m, A, B . . . C; K=u, v . . .) R, S . . . T, A, B . . . C(. . . w) are computed as explained in (6.7)-(6.13), we have the new basis equations in the explicit form

$$(6.14) \qquad x_j = b'_{jo} + \sum_{K = U, V \dots W} b'_{jK}\, x_K \qquad \begin{pmatrix} j = 1, 2 \dots) u, v \dots w, \\ r, s \dots t (\dots n+m, A, B \dots C \end{pmatrix}$$

when (6.4) is fulfilled.

The formulæ (6.14) are identities that hold for any values of the variables inside or outside the admissible region, or on its boundary provided only the determinant (6.5) is different from zero and (6.4) is fulfilled.

After having performed one freedom truncation and obtained the corresponding new basis equations (6.14), we may proceed to a new freedom truncation and so on. The whole process will appear as a gradual chopping off of variables from the basis set.

It is also possible to proceed in a different way. We can maintain all the time the same basis set and reduce the number of degrees of freedom by means of side conditions. This is explained in full in [5].

7. *How to get into the admissible region. A method of solving linear inequalities*

In our work on linear programming problems at the Oslo Institute, we have found it convenient to carry the search for an optimum through by movements in the *interior* of the admissible region. We must therefore first find a point in the interior, that is, we must find a point where all the variables

are effectively positive, not zero. From a formal viewpoint this is the same as to find a solution of a set of linear inequalities. Indeed, by (2.3) the problem can be formulated as that of finding such positive values of the variables $x_u, x_v \ldots x_w$ as will satisfy

$$(7.1) \qquad b_{jo} + \sum_{k=u,v\ldots w} b_{jk}\, x_k > 0 \quad (j=1, 2\ldots)\ u, v \ldots w\ (\ldots n+m).$$

We begin by dividing the problem in two separate problems. The first is to determine the *direction* of the movement from an initial point. The second is to determine the *length* of the movement.

Determination of the direction numbers

One way to proceed is to start from *any* point, i. e. a point where any of the variables may be positive, negative or zero, for instance conventionally the point where all the basis variables are zero. From this point we move towards a direction point determined by the completely incremental method of Section 5. The variables that are to be the selected ones in this method are determined as follows.

We determine the *deference order* of the variables, that is, first we take the strongest negative variable, i. e. the variable that is negative and with the largest absolute value, second we take the next to strongest negative variable, and so on.

Conventionally we may try to normalize the variables for instance by dividing (2.3) through by

$$(7.2) \qquad \sqrt{b_{ju}^2 + b_{jv}^2 + \ldots + b_{jw}^2}$$

but in practice hardly anything will be gained by doing so. As the work progresses, there is, however, another very useful form of normalization that will become apparent: It will become more or less clear how large a change we can expect in the various variables and the absolute size of the variables can be judged in relation to the order of magnitude of the expected change.

When all the negative variables have been exhausted, we continue with the variables that are zero, determining if necessary the order amongst them by random drawing. When also the zero variables are exhausted, we continue with the positive variables in ascending order.

When the deference order has been determined, we start from the beginning of the list and count a number of variables equal to the number of degrees of freedom in the problem. If in the process of admitting the variables one by one, we find a variable whose boundary vector happens to be lineraly dependent on the boundary vectors of the variables already admitted (which can often be seen by an inspection of rows and columns of the matrix b_{jk}), this variable is simply *skipped* and we proceed to add new variables to the set. When a final set of n variables is obtained, we use (5.5) putting

$$(7.3) \qquad d_i = \begin{cases} -x_i^o & \text{if } x_i^o < 0 \\ 0 & \text{if } x_i^o \geqq 0. \end{cases}$$

In this way a direction is determined.

Another method is to proceed only through the negative variables — or possibly through the non positive ones — and use the moment method of Section 5. This will involve the computation of moments and some extra thinking in adapting the subsequent steps. It is easier and more mechanical to use the method (7.3). Particularly if the capacity of the automatic computer is large enough to handle the system (5.5) in one stroke, much is to be said for using this method. Whichever method is used, it is a great advantage that only a one way solution, not a complete inversion, is needed.

When the direction of the movement has been fixed, the *sense* (positive or negative) and the *length* of the movement must be decided upon. To do this we will build on a principle that may be termed the principle of the *absolute value sum*. It can be stated as follows.

As we move from the initial point along the line just determined, some of or all the variables will change. In certain points some of them will also change sign, some passing perhaps from negative to positive, others from positive to negative. In any point along the line, that is for any value of λ in (5.10), we note which of the variables are *negative,* and we consider the sum

$$(7.4) \qquad S(\lambda) = \text{sum of the absolute values of all the variables that are negative (the direction numbers being given).}$$

We have the following existence theorems[1] (for brevity we drop the superscript O on the initial x_j):

If the upper and lower of the following bounds both exists (hence one of them negative and the other positive), and we have

$$(7.5) \qquad -\sum_j d_j \left[x_j = 0,\ d_j > 0 \right] \gtreqless \sum_j d_j \left[x_j < 0,\ d_j \text{ pos,. neg. or zero} \right]$$

$$\gtreqless -\sum_j d_j \left[x_j = 0,\ d_j < 0 \right]$$

1 The big bracket indicates the conditions for the summation affix. The summation is to be extended to those and only those values for which the conditions are fulfilled. If no j exists satisfying the conditions, the bound in question should be interpreted as non-existing.

there does not exist any value of λ which makes $S(\lambda)$ less than the value $S(O)$ assumed in the initial point.

In the case where (7.5) is not fulfilled, we must have either

$$(7.6) \qquad -\Sigma d_j \left[x_j = 0,\ d_j < 0 \right] < \Sigma d_j \left[x_j < 0,\ d_j \text{ pos., neg. or zero} \right]$$

or

$$(7.7) \qquad \Sigma d_j \left[x_j < 0,\ d_j \text{ pos., neg. or zero} \right] < -\Sigma d_j \left[x_j = 0,\ d_j > 0 \right].$$

If (7.6) is fulfilled, there exists a uniquely determined positive λ (or a uniquely determined positive λ interval) which produces the smallest value which the sum $S(\lambda)$ can assume for a positive λ and the value of $S(\lambda)$ in this optimum point (in this optimum intervall) is less than the value $S(O)$ in the initial point.

If (7.7) is fulfilled, there exists a uniquely determined negative λ (or a uniquely determined negative λinterval) which produces the smallest value which the sum $S(\lambda)$ can assume for a negative λ, and the value of $S(\lambda)$ in this optimum point (in this optimum interval) is less than the value $S(O)$ in the initial point.

If both inequalities (7.6) and (7.7) are satisfied (which may happen if at least one of the bounds is non-existing), there exists a positive optimum value (a positive optimum interval) of λ as well as a negative optimum value (a negative optimum interval). Which one of these two optimum values is the best, can be verified by actually computing the two values, using the algorithms described below. The necessity of such a double computation will never occur in the case where all the variables are different from O in the initial point and we have

$$\sum_j d_j \left[x_j < 0,\ d_j \text{ pos., neg. or zero} \right] \neq 0.$$

We shall first give the algorithm for determining a positive optimum value of λ.

If the criterion (7.6) for the existence of a positive optimum value of λ is satisfied, this value (this interval) is determined by first computing the magnitude V_o defined by

$$(7.8) \qquad V_o = -\sum_j d_j \left[x_j < 0,\ d_j \text{ pos., neg. or zero} \right] - \sum_j d_j \left[x_j = 0,\ d_j < 0 \right].$$

Note that in this formula it is d_j with its correct sign, positive, negative or zero, that is entered in the summation. It is not the absolute value of d_j that is entered. When (7.6) is satisfied, the magnitude V_o will be negative.

Next we compute the coefficients

$$(7.9) \qquad \lambda_j = \left| \frac{x_j}{d_j} \right| \quad \text{for all } j \text{ where } x_j \text{ and } d_j \text{ have } \textit{opposite} \text{ signs, that is, either } x_j \text{ effectively positive and } d_j \text{ effectively negative or vice versa.}$$

All the magnitudes λ_j will be positive and finite. These magnitudes are arranged in a ranking order from the lowest to the largest. Let this ranking order be

$$(7.10) \qquad 0 < \lambda_{(1)} < \lambda_{(2)} < \lambda_{(3)} < \ldots$$

that is to say $\lambda_{(1)}$ is the smallest of the numbers (7.9), $\lambda_{(2)}$ the second smallest of the numbers (7.9) etc. As a rule there will only be one value of j corresponding to $\lambda_{(1)}$, one value of j corresponding to $\lambda_{(2)}$ etc., but in principle there is nothing to prevent several values of j to correspond to $\lambda_{(1)}$, and similarly there may be several values of j corresponding to $\lambda_{(2)}$ etc.

Let $V_{(r)}$ be the *absolute value* of the d_j that corresponds to $\lambda_{(r)}$ or, if there are several values of j, namely $j = \alpha, \beta \ldots \gamma$, that correspond to $\lambda_{(r)}$, then the *sum* of these absolute values, that is to say

$$(7.11) \qquad V_{(r)} = |d_\alpha| + |d_\beta| + \ldots + |d_\gamma| \qquad (r = 1, 2 \ldots)$$

Note that in (7.11) we take the sum of the absolute values of d_j, not as in (7.8) the sum of the d_j with their correct signs.

All the magnitudes $V_{(1)}, V_{(2)} \ldots$ will be effectively positive.

Next we compute the following sequence of numbers

$$(7.12) \qquad \begin{aligned} V_{[0]} &= V_0 && \text{(computed by (7.8))} \\ V_{[1]} &= V_0 + V_{(1)} && (V_{(1)} \text{ computed by (7.11)}) \\ V_{[2]} &= V_0 + V_{(1)} + V_{(2)} && (V_{(2)} \text{ computed by (7.11)}) \\ &\ldots \end{aligned}$$

These numbers $V_{[r]}$ can be computed *in one single operation by continuous summation* directly from the absolute values of the d_j. Subtotals are taken only for $r = 1, 2 \ldots$

The initial value of the continuous summation, namely $V_{[0]} = V_o$ is negative if (7.6) is satisfied. The numbers $V_{[0]}, V_{[1]} \ldots$ form a monotonically increasing sequence. At the latest when we get to the last number in this sequence, namely

$$(7.13) \qquad V_{[\omega]} = V_0 + V_{(1)} + V_{(2)} + \ldots + V_{(\omega)}$$

where $V_{(\omega)}$ correspond to the largest $\lambda_{(\omega)}$ of the λ-coefficients computed by (7.9), the magnitude $V_{[\omega]}$ must have become non-negative. Let r be the first ranking number such that

$$(7.14) \qquad V_{[r]} \geqq 0.$$

Then if $V_{[r]} > O$, $\lambda = \lambda_{(r)}$ will be the optimum value amongst the positive λ values. That is to say, in the point $\lambda = \lambda_{(r)}$ the absolute value-sum $S(\lambda)$ assumes the smallest value which it can assume for any positive

λ. For no other positive λ is $S(\lambda)$ as small as it is in the point $\lambda = \lambda_{(r)}$ and this value $S(\lambda_{(r)})$ is less than $S(O)$. The optimum value $S(\lambda_{(r)})$ can be equal to O, in which case we have reached the admissible region, or it may happen that $S(\lambda_{(r)})$ is still positive, in which case we have not yet reached the admissible region.

If $V_{[r]} = O$, $S(\lambda)$ will maintain a constant value in the whole interval from $\lambda_{(r)}$ inclusive, to $\lambda_{(r+1)}$. And this value of $S(\lambda)$ is less than the values which can be produced by any other non-negative λ and it is actually less than $S(O)$.

The algorithm for determining a negative optimum value of λ can be given in a similar way. If the criterion (7.7) for the existence a negative optimum value (a negative optimum interval) of λ is satisfied, this λ (this interval) can be determined by first computing the magnitude $\bar{V}_o$ defined by

$$(7.15) \qquad \bar{V}_o = \Sigma d_j \left[x_j < 0,\ d_j \text{ pos., neg. or zero} \right] + \Sigma d_j \left[x_j = 0,\ d_j > 0 \right]$$

When (7.7) is satisfied, $\bar{V}_o$ must be negative.

Next we compute the coefficients

$$(7.16) \qquad \bar{\lambda}_j = \left| \frac{x_j}{d_j} \right| \quad \text{for all j where } x_j \text{ and } d_j \text{ have the } \textit{same} \text{ sign, that is to say, are either both effectively positive or both effectively negative.}$$

All the numbers (7.16) will be positive and finite. These numbers are now arranged in a ranking order from the smallest to the largest. Let this ranking order be

$$(7.17) \qquad 0 < \bar{\lambda}_{(1)} < \bar{\lambda}_{(2)} < \bar{\lambda}_{(3)} \ldots$$

In a similar way as in the case of a positive optimum value of λ, we now define

$$(7.18) \qquad \bar{V}_{(s)} = |d_\alpha| + |d_\beta| + \ldots + |d_\gamma| \qquad (s = 1, 2 \ldots),$$

where $j = \alpha, \beta \ldots \gamma$ are the numbers of the variables that by (7.16) correspond to $\bar{\lambda}_{(s)}$. All these numbers $\bar{V}_{(s)}$ are effectively positive.

Further we compute by a continuous summation

$$(7.19) \qquad \bar{V}_{[s]} = \bar{V}_o + \bar{V}_{(1)} + \bar{V}_{(2)} + \ldots + \bar{V}_{(s)} \qquad (s = 0, 1 \ldots)$$

The first of these numbers is negative when (7.17) is satisfied. The numbers $\bar{V}_{[0]}, \bar{V}_{[1]}, \ldots$ form a monotonically increasing sequence. At the latest when we get to the last number in the sequence, namely

$$(7.20) \qquad \bar{V}_{[\bar{\omega}]} = \bar{V}_o + \bar{V}_{(1)} + \bar{V}_{(2)} + \ldots + \bar{V}_{(\bar{\omega})}$$

where $\bar{V}_{(\bar{\omega})}$ correspond to the largest $\bar{\lambda}_{(\bar{\omega})}$ of the λ quotients computed by (7.16), $\bar{V}_{[\bar{\omega}]}$ must have become non-negative. Let s be the first ranking number such that

$$(7.21) \qquad \bar{V}_{[s]} \geqq O.$$

We can then draw conclusions analogous to those that were formulated in connection with (7.14).

The positive optimum value $\bar{\lambda}$ which is computed in the above way, has to be inserted with opposite sign in (5.10) in order to determine the new point x_j', that is to say, we have

$$(7.22) \qquad x_j' = x_j - \bar{\lambda} d_j \qquad (j = 1, 2 \ldots n + m)$$

The proofs are given in Section (6) in [2].

The process is repeated round after round until the admissible region is reached. If this takes place along a direction lying in the boundary (i. e. with at least one of the variables equal to zero) and we want to get effectively into the interior, we may often be able to do so by a simple inspection of the figures and a tentative modification. If this does not work quickly, we may perform one more round, inserting non-positive values for the d_i for those variables which we want to make positive. Or we may use the moment method, including in the selected set only those variables which we want to make positive.

I have not proved but believe something like the following proposition to be true:

Prop. (7.23). The necessary and sufficient condition for an empty admissible region (i. e. for the case where there exists no points giving simultaneously all the variables non-negative) is that the exceptional case (7.5) is reached after a finite number of rounds of the algorithm described above.

*

In a working team at the Indian Statistical Institute in the winter of 1954—55, one of my associates (I don't remember whom of them it was) suggested that when the linear programming problem is transformed into a problem of solving a set of linear inequalities (which is possible by a theorem of J. von Neumann), we may solve the whole linear programming problem by the $S(\lambda)$ method. I don't know how fruitful this idea is. I suspect it will depend on how large the number of degrees of freedom in the optimum set is. The pointset that in the original formulation was the optimum pointset (possibly only a single point) will appear as the admissible region in the transformed problem. In order to apply the $S(\lambda)$ method, we do not need to know beforehand the dimensionality of the pointset that forms the admissible region.

8. The logarithmic potential method for guessing about limiting variables

Suppose that we are in a point in the interior of the admissible region. If from this initial point we could make a move which would give us some basis for *guessing* about optimum variables, i. e. variables that will be zero in an optimum point, we may through one of the methods of Section 6 bring these variables to zero *and keep them there,* starting a new round with the reduced number of degrees of freedom which follows from the fact that some variables have been put equal to zero. The new round would have to start from the beginning, that is, by seeking our way into the admissible region of the new truncated space. But if the process of seeking into the admissible region is worked out in a highly mechanized way along the lines of Section 7, this computational difficulty will not be insuperable.

If instead we work by the moment method — slightly touched upon at the end of Section 6 — we do not need to work our way again and again into the admissible region. In certain other respects, however, the computational costs involved are high in the moment method, so after all, under favourable circumstances the truncation method of Section 7 will be the most economical.

Since we possess criteria for the case where an optimum point has been reached — compare Section 10 — and we can work out correctional procedures to be applied when we have reached a corner which turns out not to be optimal, the method suggested based on guesses about optimum variables will be a sound one, when conditions are such that it can be followed with a reasonable computational cost. Examples have shown that such cases exist. The great advantage of this approach is that we may under favourable circumstances freedom truncate *a large number of variables in each round* and thus approach the optimum by leaps and bounds, instead of handling one variable at a time. This is of fundamental importance when the number of variables is great.

We shall now indicate how the logarithmic potential method for guessing about optimum variables may be applied when we start from an initial point x_k° (k=u, v . . . w) in the interior of the admissible region, i. e. where all $x_j^\circ > O$ (j=1, 2 . . . n + m).

We define the potential

$$(8.1) \qquad \Omega = \Omega(x_u, x_v \ldots x_w) = \sum_{j=1,2\ldots n+m} \log x_j$$

where x_j for j=1, 2 . . .) u, v . . . x_w (. . . n + m are defined as functions of x_u, x_v . . . x_n by (2.3). In the interior of the admissible region the potential is continuous and with continuous partial derivatives of all orders, but

any point on the boundary is a singular point where the potential tends towards minus infinity. As we wander around in the admissible region the potential may therefore be used as a sort of radar which will tell us to steer away from the boundary.

The gradient components of the potential are

$$(8.2) \qquad \Omega_k = \frac{\partial \Omega}{\partial x_k} \qquad (k = u, v \ldots w)$$

and their explicit expressions in terms of the variables are

$$(8.3) \qquad \Omega_k = \sum_{j=1}^{n+m} \frac{b_{jk}}{x_j} \text{ that is } \Omega_k = \frac{1}{x_k} + \sum_{j=1,2\ldots)\,u,v\ldots w\,(\ldots n+m} \frac{b_{jk}}{x_j} \quad (k = u, v \ldots w)$$

The gradient (8.3) together with the preference vector, i e. the vector whose components are p_k (k=u, v . . .w) define two different directions in which it is in a certain sense desirable to go. In order to increase the preference function we should go in the direction p_k, but in order to steer away from the boundary, we should go in the direction Ω_k. The optimal solution consists in a wise compromise between these two directions, much in the same way as the optimal solution of the production policy of an enterprise consists in a wise compromise between a movement in the direction (in the space of the factors of production) in which the product increases most steeply and a movement in the direction in which the cost declines the most steeply.

We express this by saying that we want to move in the compromise direction

$$(8.4) \qquad d_k = p_k + c\Omega_k \qquad (k = u, v \ldots w)$$

where c is a constant to be disposed of.

To impose (8.4) where c is a constant to which we may attribute any value between — oo and + oo is the same as to say that the direction of movement from the initial point shall be in the twodimensional manifold that passes through the initial point and is unfolded by the preference vector and the gradient of the potential. The problem is to determine c in an optimal way.

First suppose that we have chosen a fixed value for c. The directional numbers d_k for the basis variables are then given by (8.4) — where we now insert the initial point values for Ω_k as computed by (8.3) —. Hence the directional numbers d_j for all the variables are given by (5.3), and the values of all the variables under a movement in this direction are given by (5.10). Written out explicitly in terms of c this becomes

$$(8.5) \qquad x_j = x_j^\circ + \lambda(p_j + c\Omega_j) \qquad (j = 1, 2 \ldots n + m)$$

where

$$(8.6) \qquad p_j = \sum_{k=u,v\ldots w} b_{jk}\, p_k \text{ and } \Omega_j = \sum_{k=u,v\ldots w} b_{jk}\, \Omega_k \quad (j = 1, 2 \ldots n + m)$$

The value of the preference function under this movement is

$$f = f(\lambda) = f^\circ + \lambda (P + cM) \tag{8.7}$$

where f° is the value of the preference function in the initial point and

$$P = \sum_{k=u,v\ldots w} p^2 \quad \text{and} \quad M = \sum_{k=u,v\ldots w} p_k \Omega_k \tag{8.8}$$

From (8.7) follows that f is a linear function of λ along the straight line considered. If we want to increase the preference function, we should let λ have the same sign as $P + cM$, i. e.

$$\text{sgn. } \lambda = \text{sgn. } (P + cM) \tag{8.9}$$

and we should make the absolute value of λ as large as is compatible with the conditions that we shall stay within the admissible region.

When the sense of the movement is fixed as indicated in (8.9), we have

$$f = f(\lambda) = f^\circ + |\lambda| \cdot |P + cM| \quad \begin{matrix}\text{(when the sign of } \lambda \text{ is} \\ \text{determined by (8.9))}\end{matrix} \tag{8.10}$$

If the initial point x_k° is in the interior of the admissible region, there exist at least some small interval of $|\lambda|$ where all the variables are still effectively positive, and consequently the point x_k still inside the admissible region. For such values of $|\lambda|$ the values of the preferense function will by (8.10) have increased. As $|\lambda|$ increases we will — if the admissible region is finite, i. e. entirely closed — sooner or later reach a *breaking out point,* that is to say, a point where at least one of the variables reaches zero and would become negative if we proceeded further.[1] The value of $|\lambda|$ for which this happens will obviously depend on c and consequently the value of the preference function in the breaking out point is a function of c. A natural principle for choosing the value of c will be to do it in such a way as to maximize the value of the preference function in the breaking out point. This will be our principle for choosing c.

This is the same as to *minimize* the function F(c) defined by

$$\frac{1}{f' - f^\circ} = F(c) = \frac{N(c)}{|P + cM|} \tag{8.11}$$

$$N(c) = \text{"Numerator"} = \max_j \left| \frac{p_j + c\Omega_j}{x_j^\circ} \right| \tag{8.12}$$

$$\left| \begin{matrix} p_j + c\Omega_j < 0 \text{ when } P + cM > 0 \\ p_j + c\Omega_j > 0 \text{ when } P + cM < 0 \end{matrix} \right.$$

f' designating the value of the preference function in the breaking out point. In other words we have to look for the solution of the minimax problem

$$\min_c \max_j \frac{\left| \frac{p_j + c\Omega_j}{x_j^\circ} \right|}{|P + cM|} \quad \left| \begin{matrix} p_j + c\Omega_j < 0 \text{ when } P + cM > 0 \\ p_j + c\Omega_j > 0 \text{ when } P + cM < 0 \end{matrix} \right. \tag{8.13}$$

1 The case where the processs indicated does not lead to a breaking out point because of an open admissible region, is trivial. In this case the preference function can be rendered arbitrarily great.

The broken line whose abscissa is c and ordinate N(c) is always convex upwards. This helps greatly. The optimum abscissa c_{opt} which solves the minimax problem (8.13), i. e. which produces the minimum of (8.11) can be determined by the following algorithm.

As c varies from —oo to +oo, we consider separately the branch of the broken (c, N) line (8.12) where $P + cM > O$ and the branch where $P + cM < O$. On each such branch we determine the point c_{opt} that furnishes the minimum of the function (8.11). On each branch the optimum is determined by a series of approximations starting from the initial value

$$c^{(1)} = \begin{cases} \frac{c_o}{2} & \text{on the branch where } P + cM > 0 \\ \frac{3c_o}{2} & \text{on the branch where } P + cM < 0 \end{cases} \tag{8.14}$$

the separating abscissa c_o for the two branches being defined by

$$c_o = \frac{P}{-M}. \tag{8.15}$$

The consecutive iteration values $c^{(2)}$ $c^{(3)}$ ... are then computed by

$$c^{(k+1)} = c^{(k)} \pm \left[c_o + \left(\frac{p_j}{\Omega_j} \right)_{j = j_{max}(c^{(k)})} \right] \quad (k = 1, 2 \ldots) \tag{8.16}$$

where

(8.17) $j_{max}(c)$ is the value of j that maximizes (8.12) for a given c.

The sign in (8.16) is chosen according to the following rule:

Rule (8.18). If the straight line $y = p_j + c\Omega_j$ determined by $j = j_{max}(c^{(k)})$ is such that along this line $|\gamma|$ changes in the *opposite* direction of $|c|$, then $|c|$ shall be increased, i. e. the sign in (8.16) shall be chosen so that we produce $|c^{(k+1)}| > |c^{(k)}|$. If $|\gamma|$ changes in the *same* direction as $|c|$ along the line determined by $j = j_{max}(c^{(k)})$, there are two cases: $|c|$ shall be *increased,* i. e. $|c^{(k+1)}|$ shall be made $> |c^{(k)}|$ if $\left|\frac{p_j}{\Omega_j}\right| < |c_o|$, but $|c|$ shall be *decreased,* i. e. $|c^{(k+1)}|$ be made $< |c^{(k)}|$ if $\left|\frac{p_j}{\Omega_j}\right| > |c_o|$.

Once the maximum of (8.12) is found for a given $c^{(k)}$, the affix $j_{max}(c^{(k)})$ is known and hence immediately $c^{(k+1)}$ by (8.16). This gives the starting point for a new iteration round of the type (8.16)—(8.18).

As we approach the optimum value c_{opt} the approximation values may begin to oscillate. The optimum must then be situated between the limits obtained. In most cases this gives a sufficiently close approximation, but if need be, we may single out the two values of j that determine the optimum point exactly.

Thus, everything hinges on *a rapid determination of the maximum to the* right in (8.12) for a given value $c^{(k)}$ This problem is well adapted for an electronic computor or a punched card outfit. One does not need a large storage capacity as all items that fall below the largest one found so far, can be discarded and only the leader retained.

When the optimum c_{opt} is determined for both branches, that one of the two c_{opt} values are taken, which produces the smallest value of the function (8.11). The two values in question are computed simply by inserting the two optimum values found for c.

When the final c_{opt} is determined, we consider the situation in the breaking out point. We compute also the values of $|\lambda|$ for some of the variables which we *would have hit if we had continued beyond the breaking out point.* To find out which variables these are, we compute systematically all the values

$$(8.19) \qquad |\lambda| = \frac{x_j^o}{-d_j} \Bigg| \; d_j < 0$$

and arrange them in ascending order. There will always be at least *two* j that give the smallest of the magnitudes (8.19) (because of the optimum determination of c_{opt} which we have performed), and these two j indicate the variables whose vanishing marked the boundary of the admissible region. The ranking order of the variables as computed by the magnitudes (8.19) we may call the *priority order* determined with respect to the breaking out point considered. This priority order we take as the principle by which to guess about optimum variables, i. e. about variables which are likely to be zero in an optimum point.[1]

The question arises of *how many* candidates, i. e. how many variables to include in our guess.

Empirically we have found that it pays fairly well to include at each round additional candidates equal in number to *the square root of the number of degrees of freedom with which we arrive in the breaking-out-point in question.*

Since this rule will usually give fractional numbers, we use a standard procedure for the rounding off to integers.

We first compute by the following recurrence formulae

$$(8.20) \qquad N_t = N_{t-1} - \sqrt{N_{t-1}}$$

the number of degrees of freedom N_t which should theoretically be retained

[1] The idea of using values of the $|\lambda|_j$ for ranging the candidates in a priority order was suggested some years ago by Mrs. Inger Haugstad, Chief computor in the University Institute of Economics, Oslo. Later more refined principles have been tried successfully but they involve higher computational costs.

in round No. t, the initial value being $N_o = n$. These numbers N_t are then rounded off to the nearest integer

$$(8.22) \qquad n_t = \text{nearest integer to } N_t, \text{ and}$$

n_{t-1}—n_t is taken as the number of *new* candidates to include in our optimality guess before we start on round No. t. Or otherwise expressed: When going from round No. t—1 to round No. t, we *reduce the number of degrees of freedom* by

$$(8.22) \qquad n_{t-1} - n_t.$$

9. The moment method for guessing about limiting variables

We may also use the moment method for guessing about limiting variables. This problem is discussed in [5] and [7].

10. Criteria for optimality

Before working out computational rules for testing a point for optimality, it will be well to state certain general principles.

(10.1) *Sufficient condition for optimality, general formulation* (any point on the boundary).

A given point is optimal if there exist n linearly independent variables that can be divided in two subsets, one of ν and the other of $(n-\nu)$ variables $(0 \leqq \nu \leqq n)$ such that:

(a) The preference prices[1] of the ν variables are negative and the preference prices of the $(n-\nu)$ variables are zero. (The condition (a) is a property independent of the point considered).

(b) In the point considered the ν variables are zero and the $(n-\nu)$ variables are non-negative.

(c) All the other m variables are non-negative.

Proof. Let the $n-\nu$ variables be Nos. $\varkappa = \alpha, \beta \ldots \gamma$ and let the ν variables Nos. $k = u, v \ldots) \alpha, \beta \ldots \gamma (\ldots w$. By hypothesis the preference function can be written in the form

$$(d) \qquad f = p_o + \sum_{k=u,v\ldots)\alpha,\beta\ldots\gamma(\ldots w} p_k x_k$$

where all the p under the summation sign are negative. This formula applies in any point x_k $(k = u, v \ldots w)$ inside or outside of the admissible region or on its boundary. Let x_k' $(k = u, v \ldots w)$ be any other point inside or outside the admissible region or on its boundary, and let f' be the corresponding value of the preference function. We then have

[1] By preference prices is here meant the coefficients which emerge when the $\nu + (n - \nu) = n$ variables considered are taken as basis set.

(e) $$f^1 - f = \sum_{k = u, v \ldots)a, \beta \ldots \gamma(\ldots w} p_k (x_k' - x_k)$$

If all the p_k in this summation are effectively negative, it is impossible to make the difference f'-f positive without making at least one of the differences $(x_k' - x_k)$ negative. This, however, is impossible if all $x_k = 0$ and none of the x_k can be made negative. Hence, if also (c) is fulfilled, the point considered must be optimal. The essence of the proof is that *no positive* preference prices must occur and for those preference prices that are negative, the corresponding values of the variables must be zero.

(10.2) *Sufficient condition for optimality, special formulation* (a corner on the boundary).

A given point is optimal if there exists n linearly independent variables such that:

(a) The preference prices of these n variables are non-positive. (This is a condition independent of the point considered).

(b) In the point considered the n variables in question are zero.

(c) All the other m variables are non-negative.

The special formulation is obtained from the general simply by considering the case where not only the ν but also the $(n\text{-}\nu)$ variables are actually zero.

(10.3) *Necessary condition for optimality, general formulation* (any point on the boundary).

A point that is optimal must satisfy (10.1) for some value of ν. If we disregard the case where the total admissible region is optimal, we can even say that it must satisfy (10.1) with $1 \leqq \nu$.

The proof will not be given here. It is discussed in [2].

(10.4) *Necessary condition for optimality, special formulation* (a corner).

In a point where all the basis variables are zero and all the dependent variables effectively positive (a simply determined corner) it is not only *sufficient* for optimality that all the preference prices p_k (k=u, v ... w) are non-positive — as follows from (10.2) — but this latter condition is also *necessary*.

If at least one of the dependent variables is zero, necessity does not follow. In other words, a point in the admissible region where all the basis variables are zero and also at least one dependent variable is zero (a multiply determined corner)[1] may be optimal even though not all the preference prices are non-positive. But by (10.2) if it is optimal, it must be possible to define the *same* corner by means of other basis variables in such a way that (10.1) can *be applied.*

(10.5) *Degrees of freedom of the optimal region.*

If the equations can be transformed to a basis form where $n\text{-}\nu$ of the preference prices are zero — compare (10.1) — and the other ν preference prices are negative, not zero, every point that is situated in the admissible region (i e. has all the n + m variables non-negative) and belongs to the $(n\text{-}\nu)$ dimensional linear manifold generated by putting the ν variables equal to zero and letting the $n\text{-}\nu$ variables perform an independent variation, is optimal. Over the linear manifold in question we have $f = p_o$ and

$$(10.6) \qquad x_j = b_{jo} + \sum_{\varkappa = \alpha, \beta, \ldots \gamma} b_{j\varkappa} x_\varkappa \qquad (j = 1, 2 \ldots n + m)$$

The rule (10.5) follows directly from (10.1).

In [7] these optimality principles are applied to the results of some of the computational procedures discussed above.

*

If the total optimality criterion does not turn out in the affirmative, different procedures may be followed. One method is to move a small distance into the interior of the admissible region by one of the directional techniques, and make a fresh start. If the move into the interior is made in a reasonable way, one will get a new starting point from where it is possible to make a much more rapid progress than from the original starting point. As another alternative, we may — if necessary by the method of Section 7 — make sure that we get a corner of the boundary and if this point does not give a positive answer to the total tests (10.10) — (10.13), we may finish up by some rounds of the simplex method. As the bulk of the variables will already have been brought into order, it is to be expected that this final work will go through rapidly.

More precisely expressed: It is fair to assume that the logarithmic potential method will have brought us to a part of the boundary where a corner (a vertex) which we have reached is a *neighbour* of an optimal corner, or nearly a neighbour. This means that the corner in question is separated from an optimal corner by one edge only or at most by a chain consisting of a small number of edges. If this is the case, the simplex method ought to carry us rapidly from a corner determined by the logarithmic potential method through a final phase towards an optimal point.

[1] When a manifold on the boundary (in the special case $\nu = n$, a corner) is multiply determined, it is customary to speak of "degeneracy". This term is not a happy one. It seems more appropriate and suggestive to speak of "a multiply determined" manifold.