PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A
MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Analysis of Complex Systems

G. W. T. White and M. D. Simmons

Phil. Trans. R. Soc. Lond. A 1977 **287**, 405-423
doi: 10.1098/rsta.1977.0151

Phil. Trans. R. Soc. Lond. A. **287**, 405–423 (1977) [405]
Printed in Great Britain

Analysis of complex systems

By G. W. T. White and M. D. Simmons
University Engineering Department, Control and Management Systems Division,
University of Cambridge, Cambridge

It is argued that a complex system of the kind arising in industry is intrinsically structured: that is to say it may be regarded as an interconnected assembly of subsystems. By exploiting this structure it is possible to decompose the problem of controlling a complex system into a series of interlinked subproblems of manageable size. Each subproblem can be solved largely independently of the others, with the interconnections between the subproblems being accounted for by some kind of coordination procedure. Such an approach to complex systems analysis leads naturally to a study of decentralized and hierarchical control methods and as there is no doubt that industrial management has a hierarchical structure it is important to assess the usefulness of the techniques of hierarchical systems theory in the management of complex industrial systems. In cases where the overall management problem can be posed as a linear or non-linear programming problem there is a good body of theory to support the decentralized approach and the basic concepts of this theory are illustrated in this paper by consideration of decentralized optimization of an interconnected production system. It is shown that effective optimal coordination of such a system is difficult to achieve if the subsystems are nonlinear but that satisfactory coordination procedures may be devised if proper account is taken of the slackness which usually exists in the interconnections between real systems. The need to assess the usefulness of these ideas in a real industrial environment is stressed.

1. Introduction

From an operations research viewpoint, effective management and control of large systems is difficult because, among other things, the task of formulating adequate models and objectives is very complex, and because sheer size makes the techniques for using the models and computing the optimal values of the objectives very slow or even impractical. Nevertheless the rewards for success are high, and during the last fifteen years considerable effort has been devoted to the analysis of large-scale problems and to the synthesis of effective techniques for solving these problems – see, for example, Himmelblau (1973) or the Proceedings of the I.F.A.C. Symposium on Large Scale Systems Theory and Applications, Udine, Italy (1976). As an area of research and application this challenging work has attracted the attention of not only the management scientist but also the control engineer, the applied mathematician and other specialists, and in the literature on large-scale systems we find contributions from many disciplines. A study of this literature reveals the underlying belief that the fundamental characteristic of a large, complex system is that it is not an amorphous aggregate but a purposefully interlinked assembly of units or subsystems. Likewise, it is the thesis of this paper that complex industrial systems are structured, and knowledge of this structure should be exploited in the modelling, control and management of the complex. Any global problem requiring solution for the whole complex should be broken down into a set of subproblems, one associated with each subsystem; these subproblems are then solved independently of each other but under the influence of a coordinator whose task it is to account for the interconnections and conflicts between the subsystems. Thus we are concerned with the concepts of decomposition and coordination, of hierarchical control and of decentralization.

There is a lack of consensus about the precise meaning of the terms 'decentralized' and 'hierarchical'. In the literature, and to some extent in this paper, the two terms are often used synonymously. Nevertheless it is sometimes desirable to distinguish between the concept of a decentralized system and that of a hierarchical system, and there is growing support (Sandell, Varaiya & Athans 1976; Wilson 1977) for the following view. A centralized system is one in which all the system information is available centrally, and in which all the system variables may be controlled directly from the centre. Conversely in a decentralized system there is more than one controller, each of which has knowledge of only strict subsets of the system information, and is able to manipulate only strict subsets of the system variables. A decentralized system is hierarchical only if the information sets of some controllers depend directly on the action of other controllers, thus establishing a priority of intervention of some controllers over others. However, a hierarchical system is not decentralized if any controller may operate directly on all the system variables.

These concepts, and some of the problem solving techniques resulting from them, are applicable to many aspects of management science including information structures and decision making (Athans 1974; Ho & Chu 1974; Bailey 1976), scheduling (Drew 1975) and control (Findeisen 1974*a*). In much of the work in these areas it is explicitly or implicitly assumed that the problems which arise can be posed as constrained, nonlinear optimization problems such as the minimization of a cost functional or the maximization of some profit measure. Admittedly in many practical cases the problem solutions may be sought by methods other than optimization, but it is believed that the information structures needed to solve the problems by decentralized hierarchical optimization will be similar to the structures that are needed for any other decentralized hierarchical solution method. Thus it is the purpose of this paper to assess the various decomposition and coordination techniques whereby optimization problems may be solved in a decentralized hierarchical fashion. The optimization problem can arise in many forms. The variables may be subject to stochastic constraints and the objective function may be an expectation, or the whole problem may be deterministic. The constraints may be differential equations and the objective function may be an integral with respect to time, i.e. the problem may be a dynamic one, or the problem may be static. However, the various strategies for decomposing an optimization do not necessarily depend upon the type of problem, and so for the purposes of introducing the basic decomposition and coordination techniques and delineating the associated structures of information flow, consideration of the deterministic static problem is sufficient and in this paper attention is restricted to this type of problem; it is not only the simplest type of problem, and therefore the best suited to illustrate the methodology, but it is also the type of problem whose solution has the greatest potential for application.

The layout of the paper is as follows. Section 2 contains a discussion of the motivation for considering decentralized optimization in an industrial environment, and in §3 a brief review of the basic theory is given. In §4 the relevance of this theory to real problems is discussed and emphasis is given to the rôle which approximation plays in formulating and solving practical problems. This leads to consideration in §5 of more recent methods of decentralized optimization designed to solve problems which are not rigidly constrained.

A summary of the conclusions is presented in §6.

2. Motivation for decomposition

Before reviewing the theory of some of the basic methods for decomposing nonlinear programming problems it is worth while questioning why decomposition should be considered in the first place. The current generation of computers is very large and fast, and great improvements in nonlinear programming algorithms have taken place. The solution of optimization problems involving about 100 variables and 100 nonlinear constraints in reasonable time is now often possible. While this does not stand comparison with the very powerful linear programming capability available today, there are many nonlinear programming problems arising in industry which can be perfectly satisfactorily solved centrally (with a single level algorithm). Thus it is not always true to say that decomposition is necessary for the solution of large nonlinear optimization problems arising in industry. Nor is it true to say that decomposition will generally lead to reduced computation times; frequently the converse is true. Admittedly the decomposition of a large mathematical programming problem can ease the difficulties of limited computer storage, because the subsystem problems provide a natural modular framework for overlaying what could otherwise be much too large a program and data set, and if the subsystem optimizations can be solved *in parallel* on separate but linked computers then it is more likely that savings in computer time can be made. However, these are almost incidental benefits or natural consequences of a deeper motivation for adopting a decentralized approach.

It is important to remember that the optimization problems we are studying arise from an attempt to manage and control a complex system – frequently a large complex system. Optimization is not an end in itself but a means to an end, and it must be subservient to the management objectives. If the management tasks are organized on a hierarchical decentralized basis, as they usually are, and if optimization is part of the management approach, then hierarchical decentralized optimization arises naturally, and it is vital that we should understand the properties of a set of interlinked optimizers arising in this way. This last point constitutes the real reason for the study of decentralized optimization: because hierarchical optimization is a natural corollary of hierarchical management, which by observation is the type of structure used to control most complex systems, it is essential to study the conditions under which hierarchical optimization is effective. By 'effective' we mean able to obtain the correct answers reliably, with reasonable computing effort and moderate data transmission requirements, and yet able to provide all of the answers which management needs in an industrial environment.

At this point it is useful to consider briefly the influences which computers have had on the management of large industrial systems, because these influences also affect our interest in decentralized optimization. Initially the powerful data handling and computational power of computers, combined with the growing development of O.R. techniques and management science, led to the idea of totally integrated management based on a central computer. For a large-scale complex system this is quite impossible. Indeed to quote from the Steel Industry Project Staff of Purdue University (1975), 'The early visualized concept of controlling the whole steel mill from one central computer system proved impractical because of the previously unconceived of complexity and magnitude of the task and the totally unappreciated difficulty of programming very large systems into a single computer.' Thus it is necessary to break down this vast computing task into a number of problems of manageable size and in doing this one arrives at a computing structure closely related to the management structure which in turn is related to the intrinsic structure of the industrial complex. Discussion of the task oriented structures that arise as a result of decomposition in large industrial complexes has been given by a number of authors (Mesarovic, Macko & Takahara 1970; Lefkowitz 1966). Wilson (1977) has described this breaking down process formally in terms of problem decomposition, whereas Cheliustkin & Lefkowitz (1975) and Findeisen & Lefkowitz (1969) have preferred a more heuristic approach and in their functional multilayer hierarchy they recognize four principal levels: the regulation or direct control level, the supervisory or optimization level, the adaption level and the self organizing level. It is now commonplace, particularly in the process industries such as steel and petrochemicals, to find a hierarchy of computers associated with these functional levels. Typically there will be on-line computers for direct control of the plants, and the set points (desired regulation levels) for these computers will be provided by supervisory computers: yet again above these there may be what can be termed a management information computer primarily concerned with assisting the commercial decision making. Optimization techniques may be used at all three levels of this computer hierarchy but such techniques are most usually exploited at the supervisory level. In the case where several supervisory computers are used, optimization can be truly decentralized and the optimization of each plant or group of plants can be carried out in the computer associated with that plant or group.

These ideas are well illustrated by the approach to production planning and plant control used by many petrochemical industries (Dyer 1976). As discussed above, the size of a modern petrochemical complex makes it impractical to attempt a single level on-line optimization of the entire business and instead a simplified production planning model is used off-line at regular intervals to maximize profitability by setting the desired steady state operating conditions of the various production units. At the supervisory level, on-line computers which use more detailed models may then optimize the instantaneous profit rate of each plant while maintaining the desired average operating conditions. The off-line model is designed to give the gross profit over a period of a month or more as a function of commercial and production planning decisions taken for the entire complex. Such a model must necessarily be a simplified one but even so the number of variables and constraints is so large that only a linear model can be considered and the optimization at this level is achieved by linear programming. At the supervisory level more detailed nonlinear models are required in order to represent accurately enough the response of the plants to controllers and disturbances. The optimization at this level will usually be by nonlinear programming. In this situation Dyer draws attention to the problem of ensuring that the local on-line optimizers are consistent with the overall optimal strategy for the whole complex computed by the off-line linear program. The output of the linear program contains both the desired production levels and the marginal prices for the various products and feedstocks which flow between the plants. It is by no means clear whether all this information can be used consistently by the on-line optimizers, nor how it can be used – a problem which is confounded by the differences of detail and time scale between the off-line and on-line models.

The theory of decentralized optimization is concerned with just these problems but it can give only partial answers, and successful resolution of these problems in industry depends critically upon the skill of the O.R. team who build and maintain the models and run the optimizers, and above all on the facilities provided for management supervision and intervention which are an integral part of successful planning, scheduling and optimization schemes in industry.

3. Theory of decentralized optimization

The aim of this section is to review sufficient of the basic theory of decentralized optimization to allow a discussion of the advantages and disadvantages of the technique and its associated algorithmic problems, and to provide a basis for the consideration of practical usefulness of the methods and possible extensions presented in §§4 and 5. The theory of decentralized optimization has been treated in depth in a number of authoritative works, among which the books by Lasdon (1970) and Findeisen (1974*a*) are notable. The former concentrates on large-scale linear and, to a lesser extent, nonlinear problems arising typically in an operations research environment, and gives a detailed analysis of the decomposition algorithms of Dantzig & Wolfe (1960), Benders (1962) and Rosen (1964). The latter is primarily concerned with separable optimization problems arising in multilevel control systems. The basic manipulations and strategies used in decomposing optimization problems have been reviewed by Geoffrion (1970) and more recently by Wilson (1977). The theory is not new therefore, neither can such a brief review as that given below fully explain all the variations that are possible. What follows is a discussion of the most basic methods of decomposing and coordinating nonlinear programming problems: the treatment follows closely that given by Simmons (1975) and Findeisen (1974*b*).

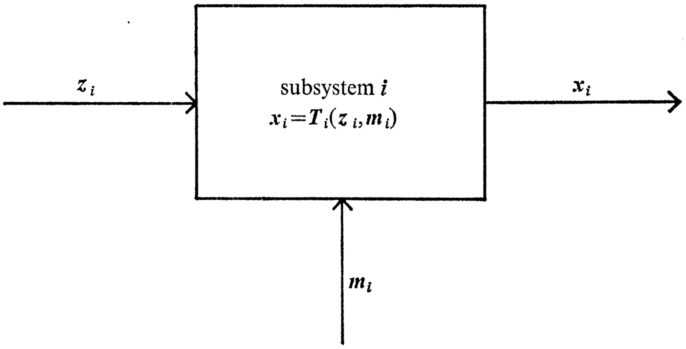

Figure 1. Subsystem *i* of an interconnected production system.

(*a*) *A mathematical programming problem for a complex production system*

To avoid difficulties of understanding arising from unnecessary mathematical abstraction, and to emphasize the relevance of the theory to real problems, we derive the mathematical programming problem, which forms the starting point for our treatment of the theory, by considering a disaggregated production system. Suppose an assembly of N subsystems (factories or process units) is interconnected by product streams. As shown in figure 1 the *i*th subsystem has a vector of inputs $\boldsymbol{z}_i$ from the other subsystems, a vector of outputs $\boldsymbol{x}_i$ to the other subsystems and a vector of controls $\boldsymbol{m}_i$. It is assumed that the algebraic equation describing the performance of the subsystem, i.e. the model of the subsystem, is

$$\boldsymbol{x}_i = \boldsymbol{T}_i(\boldsymbol{z}_i, \boldsymbol{m}_i) \tag{1}$$

and that the feasible operating region of the subsystem is defined by the vector of local constraints

$$\boldsymbol{h}_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) \leqslant \boldsymbol{0}. \tag{2}$$

The interconnections between the subsystems are simple in the sense that an output stream can be characterized by a single variable and goes directly from one subsystem to another, although stream merging and recycling are allowed; under these conditions we may write

$$G\begin{bmatrix}\boldsymbol{z}\\ \boldsymbol{x}\end{bmatrix} = \boldsymbol{0}, \tag{3}$$

where G is a matrix of ones and zeros, generally sparse and non-square, and where $\boldsymbol{z}$ and $\boldsymbol{x}$ are the composite input and output vectors

$$[\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_2^{\mathrm{T}} \dots \boldsymbol{z}_N^{\mathrm{T}}]^{\mathrm{T}} \quad \text{and} \quad [\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2^{\mathrm{T}} \dots \boldsymbol{x}_N^{\mathrm{T}}]^{\mathrm{T}}.$$

There are several ways in which the matrix G can be partitioned; for the methods described in this section it is convenient to proceed by suitably permuting the columns of G and ordering the elements of $\boldsymbol{z}$ and $\boldsymbol{x}$ to group together the inputs and outputs of each subsystem so that equation (3) may be written

$$[G_1 \vdots G_2 \vdots \dots \vdots G_N]\begin{bmatrix}\boldsymbol{z}_1\\ \boldsymbol{x}_1\\ \cdots\\ \boldsymbol{z}_2\\ \vdots\\ \boldsymbol{x}_N\end{bmatrix} = \boldsymbol{0}$$

or

$$\sum_{i=1}^{N} G_i\begin{bmatrix}\boldsymbol{z}_i\\ \boldsymbol{x}_i\end{bmatrix} = \boldsymbol{0}. \tag{4}$$

For convenience of notation it is desirable to write this as

$$\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i) = \boldsymbol{0}, \tag{5}$$

where the $\boldsymbol{g}_i$ are linear vector functions having a dimension equal to the number of rows of G, i.e. equal to the total number of product streams. It is assumed that there is a scalar objective function for each subsystem, $f_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i)$, and that the global objective function is just the sum of the subsystem objectives

$$\sum_{i=1}^{N} f_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i,\, \boldsymbol{m}_i). \tag{6}$$

Thus the global mathematical programming problem which we denote by M and which we assume to be well posed is

$$M \quad \max_{\boldsymbol{z},\boldsymbol{x},\boldsymbol{m}} \sum_{i=1}^{N} f_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i,\, \boldsymbol{m}_i) \tag{6}$$

subject to

$$\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i) = \boldsymbol{0} \tag{5}$$

$$\left.\begin{aligned}\boldsymbol{x}_i &= \boldsymbol{T}_i(\boldsymbol{z}_i,\, \boldsymbol{m}_i) \qquad (1)\\ \boldsymbol{h}_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i,\, \boldsymbol{m}_i) &\leqslant \boldsymbol{0} \qquad (2)\end{aligned}\right\}\; i = 1, 2, \dots, N.$$

For later use let us define $\mathring{\boldsymbol{z}}$, $\mathring{\boldsymbol{x}}$, $\mathring{\boldsymbol{m}}$ to be the solution of this problem and let

$$\sum_{i=1}^{N} f_i(\mathring{\boldsymbol{z}}_i,\, \mathring{\boldsymbol{x}}_i,\, \mathring{\boldsymbol{m}}_i) = F.$$

What we seek now are ways in which M may be broken down into N independent optimization problems which can be coordinated in some way to achieve the solution of the global problem. We describe below two basic methods: primal coordination and dual coordination. A study of these two methods is an essential basis for understanding the various decomposition and coordination procedures described in the literature, and provide a background for the discussions which follow in §§4 and 5.

(*b*) *Primal coordination*

In this method the solution of M is sought by maximizing the global objective function iteratively, first over the controls $\boldsymbol{m}$, then over the inputs and outputs $\boldsymbol{z}$ and $\boldsymbol{x}$, then again over $\boldsymbol{m}$ and so on until convergence. This is essentially Geoffrion's (1970) method of projection, Mesarovic *et al.*'s (1970) method of model coordination and what Kulikowski, Krus, Manczak & Straszak (1975) and Wilson (1977) call parametric decomposition. To analyse the conditions under which the method will work it is necessary to describe it more formally which we may do as follows. If the interconnection variables $\boldsymbol{z}$ and $\boldsymbol{x}$ are fixed at some value which satisfies the interconnection constraint (5), then the maximization of (6) is only over the controls $\boldsymbol{m}$, and the problem M separates into N independent maximization problems

$$P(i)\quad \max_{\boldsymbol{m}_i} f_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i,\, \boldsymbol{m}_i)$$

$$\left.\begin{aligned} \text{s.t.}\quad \boldsymbol{h}_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i,\, \boldsymbol{m}_i) &\leqslant \boldsymbol{0},\\ \boldsymbol{x}_i &= \boldsymbol{T}_i(\boldsymbol{z}_i,\, \boldsymbol{m}_i).\end{aligned}\right\} \qquad (7\mathrm{i})$$

Let the solution of these problems be denoted by $\overset{*}{\boldsymbol{m}}_i$ (more correctly by $\overset{*}{\boldsymbol{m}}_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i)$ because these solutions depend upon the parameters $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$, but the inclusion of the arguments is notationally clumsy and is therefore omitted). The value of the global objective function for these values of $\boldsymbol{m}_i$ (and the chosen values of $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$) will be

$$\sum_{i=1}^{N} f_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i,\, \overset{*}{\boldsymbol{m}}_i) = \psi(\boldsymbol{z},\, \boldsymbol{x}) \quad \text{say}$$

and the aim is to maximize this function. Thus the successive values of $\boldsymbol{z}$ and $\boldsymbol{x}$ will be determined by solution of the coordinator or master problem

$$CP\quad \max_{\boldsymbol{z},\,\boldsymbol{x}} \psi(\boldsymbol{z},\, \boldsymbol{x})$$

$$\text{and}\qquad \left.\begin{aligned} \text{s.t.}\ \sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i) &= \boldsymbol{0},\\ (\boldsymbol{z},\, \boldsymbol{x}) &\in V,\end{aligned}\right\} \qquad (8)$$

where the set V is introduced to ensure that it is possible to find solutions to the problems $P(i)$ The set V is defined as follows

$$V = \{(\boldsymbol{z},\, \boldsymbol{x})\,|\ \text{there exist } \boldsymbol{m}_i \text{ satisfying equations (7i) for all } i\}. \qquad (9)$$

The problems $P(i)$ and CP are totally interdependent. In one iteration of the algorithm the local problems $P(i)$ take the values of $\boldsymbol{z}$ and $\boldsymbol{x}$ specified by CP and hand back the optimal value of the local objective functions $f_i(\boldsymbol{z}_i,\, \boldsymbol{x}_i,\, \overset{*}{\boldsymbol{m}}_i)$; the coordinator uses these values to compute new feasible values of $\boldsymbol{z}$, $\boldsymbol{x}$ which will increase ψ, and sends these values to the sub-systems ready for the next iteration. A diagram of this information flow for a three subsystem problem is shown in figure 2.

It should be noted that because CP is constrained to work with feasible $\boldsymbol{z}$ and $\boldsymbol{x}$, i.e. values of $\boldsymbol{z}$ and $\boldsymbol{x}$ satisfying (5), the successive values of $\overset{*}{\boldsymbol{m}}_i$ generated as the iterations progress, will maintain the stream interconnections in balance, although not at their optimal values until the iterations have converged. This property, which is not possessed by the dual coordination method described below, is useful in on-line situations. Also in contrast to the later methods, primal coordination does not introduce any auxiliary variables to achieve decomposition and thus the sum of the dimensions (number of variables) of the problems *CP* and $P(i)$ is the same as the dimension of M. However, this very point is intrinsically an important source of difficulty with primal coordination – in working with the minimum number of variables the algorithm may not give sufficient degrees of freedom to permit solution of the subsystem problems. It is quite possible for a subsystem to have more inputs and outputs than controls and thus if the inputs and outputs are arbitrarily fixed it is likely that there are no feasible values of $\boldsymbol{m}_i$ which can achieve satisfaction of the local constraints (7). This is part of the reason for introducing the set V into *CP* but the simple definition of V given in (9) belies the fact that it is very difficult to compute this set at the coordinator level without centralizing all the local information (Findeisen 1974*b*), although some resolution of this difficulty is possible if all the functions $\boldsymbol{T}_i$ and $\boldsymbol{h}_i$ are linear (Kulikowski *et al.* 1975).

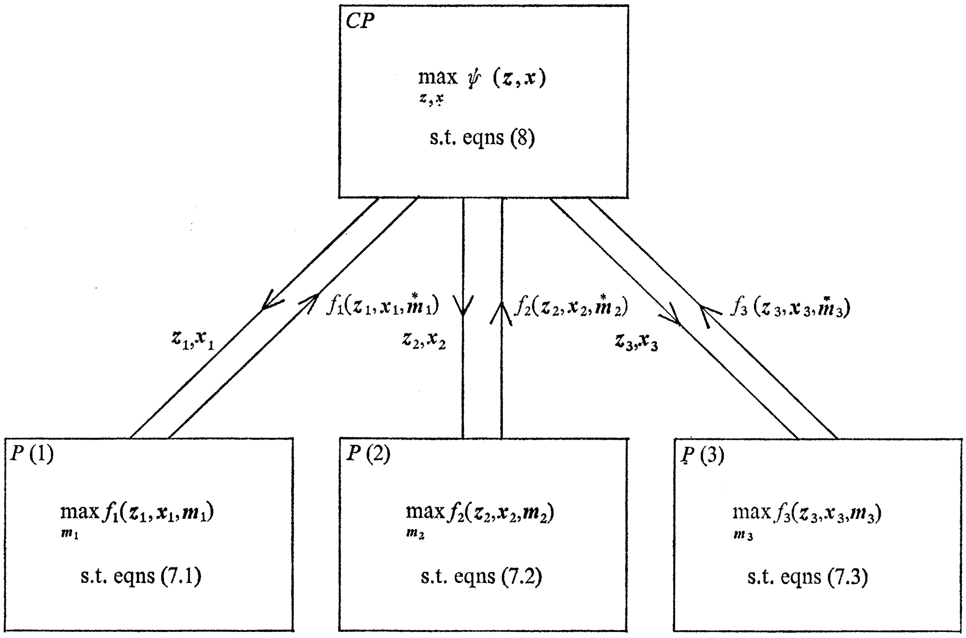
CP
$\max_{z,x} \psi(z,x)$
s.t. eqns (8)
$f_1(z_1,x_1,\overset{*}{m}_1)$
z_1,x_1
z_2,x_2
$f_2(z_2,x_2,\overset{*}{m}_2)$
z_3,x_3
$f_3(z_3,x_3,\overset{*}{m}_3)$
P (1)
$\max_{m_1} f_1(z_1,x_1,m_1)$
s.t. eqns (7.1)
P (2)
$\max_{m_2} f_2(z_2,x_2,m_2)$
s.t. eqns (7.2)
P (3)
$\max_{m_3} f_3(z_3,x_3,m_3)$
s.t. eqns (7.3)

FIGURE 2. Information flow in primal coordination.

Another source of difficulty is that, in general, the gradient of the objective function $\psi(\boldsymbol{z}, \boldsymbol{x})$ is not easy to compute and may not even exist: this is an important limitation because one iteration of the coordinator problem requires solution of N problems $P(i)$. Thus it is desirable to solve *CP* efficiently, and this is not really possible without gradient information. Even if assumptions are made that the functions f_i are strictly concave, the functions $\boldsymbol{h}_i$ are convex and differentiable and the functions $\boldsymbol{T}_i$ are linear, the function ψ may still not be differentiable but it can be shown that ψ is concave. Therefore subdifferentials of ψ exist and may be used as a basis for effective optimization algorithms although details of such algorithms for solving *CP* have been worked out only for rather special cases (Lasdon 1970).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

These difficulties with primal coordination, which stem in part from the inflexibility of equality coupling constraints, constitute just the kind of difficulties which must be overcome in industry when coordination of a complex production system is attempted by specifying the product rates and feedstock consumptions of each factory or process unit. In practice these difficulties are largely overcome by the provision of product storage thus allowing some relaxation of the interconnection constraints and in §§4 and 5 consideration is given to how this relaxation can be exploited in the context of decentralized optimization; but first an alternative method of coordination is presented.

(*c*) *Dual coordination*

Dual coordination is based on strong Lagrangian theory (Whittle 1971). The coupling constraints (5) are adjoined to the objective function (6) using a vector of Lagrange multipliers of a dimension equal to the number of elements of $\boldsymbol{g}_i$, i.e. equal to the number of interconnection streams. Thus a global Lagrangian function is defined as

$$L(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\lambda}) = \sum_{i=1}^{N} f_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) - \boldsymbol{\lambda}^{\mathrm{T}} \sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{z}_i, \boldsymbol{x}_i) \qquad (10)$$

and the maximum of this function is sought subject only to the local constraints (1) and (2); this maximization problem we denote by L.

$$L \quad \max_{\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{m}} L(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\lambda})$$

$$\text{s.t.} \quad \left.\begin{array}{r} \boldsymbol{x}_i = \boldsymbol{T}_i(\boldsymbol{z}_i, \boldsymbol{m}_i) \\ \boldsymbol{h}_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) \leqslant \boldsymbol{0} \end{array}\right\} \text{ for all } i = 1, 2, \ldots, N.$$

Let the values of the inputs, outputs and controls which solve L for any given value of $\boldsymbol{\lambda}$ be denoted by $\overset{*}{\boldsymbol{z}}$, $\overset{*}{\boldsymbol{x}}$ and $\overset{*}{\boldsymbol{m}}$. (Again these should more correctly be denoted by $\overset{*}{\boldsymbol{z}}(\boldsymbol{\lambda})$, $\overset{*}{\boldsymbol{x}}(\boldsymbol{\lambda})$, $\overset{*}{\boldsymbol{m}}(\boldsymbol{\lambda})$ because they depend upon the given value of $\boldsymbol{\lambda}$ but the functional notation is omitted for convenience.) *If* a value of $\boldsymbol{\lambda}$, say $\mathring{\boldsymbol{\lambda}}$, can be found such that the interconnection constraints are satisfied by the resulting values of $\overset{*}{\boldsymbol{z}}$ and $\overset{*}{\boldsymbol{x}}$, i.e. such that

$$\sum_{i=1}^{N} \boldsymbol{g}_i(\overset{*}{\boldsymbol{z}}_i, \overset{*}{\boldsymbol{x}}_i) = \boldsymbol{0}, \qquad (11)$$

then $$\overset{*}{\boldsymbol{z}} = \mathring{\boldsymbol{z}},\ \overset{*}{\boldsymbol{x}} = \mathring{\boldsymbol{x}},\ \overset{*}{\boldsymbol{m}} = \mathring{\boldsymbol{m}} \quad \text{and} \quad L(\overset{*}{\boldsymbol{z}}, \overset{*}{\boldsymbol{x}}, \overset{*}{\boldsymbol{m}}, \mathring{\boldsymbol{\lambda}}) = F,$$

that is to say the solution of L for $\boldsymbol{\lambda} = \mathring{\boldsymbol{\lambda}}$ solves M. For any other value of $\boldsymbol{\lambda}$ it may be shown that

$$L(\overset{*}{\boldsymbol{z}}, \overset{*}{\boldsymbol{x}}, \overset{*}{\boldsymbol{m}}, \boldsymbol{\lambda}) \geqslant F, \qquad (12)$$

i.e. the maximum value of the Lagrangian for any given value of $\boldsymbol{\lambda}$ provides an upper bound on F and this upper bound will equal F if and only if there exist a (finite) $\boldsymbol{\lambda} = \mathring{\boldsymbol{\lambda}}$, such that (11) is satisfied (Whittle 1971). It follows from this that the required value of $\boldsymbol{\lambda}$ may be sought by minimizing $L(\overset{*}{\boldsymbol{z}}, \overset{*}{\boldsymbol{x}}, \overset{*}{\boldsymbol{m}}, \boldsymbol{\lambda})$ over $\boldsymbol{\lambda}$. Thus we define

$$\phi(\boldsymbol{\lambda}) = L(\overset{*}{\boldsymbol{z}}, \overset{*}{\boldsymbol{x}}, \overset{*}{\boldsymbol{m}}, \boldsymbol{\lambda}) \qquad (13)$$

and attempt to determine $\mathring{\boldsymbol{\lambda}}$ by solution of the problem D.

$$D \quad \min_{\boldsymbol{\lambda}} \phi(\boldsymbol{\lambda}).$$

The function $\phi(\boldsymbol{\lambda})$ is called the dual function and D the dual problem (i.e. the dual of M).

The pair of interacting problems L and D now replace M in the sense that the solution of M is sought by iteratively solving L and D in succession. The point of this problem transformation is that for a given value of $\boldsymbol{\lambda}$ the Lagrangian separates; thus from (10)

$$\begin{aligned} L(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\lambda}) &= \sum_{i=1}^{N} f_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) - \boldsymbol{\lambda}^{\mathrm{T}} \sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{z}_i, \boldsymbol{x}_i), \\ &= \sum_{i=1}^{N} \{f_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}_i(\boldsymbol{z}_i, \boldsymbol{x}_i)\}, \\ &= \sum_{i=1}^{N} l_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i, \boldsymbol{\lambda}), \end{aligned}$$

where the scalar function $l_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i, \boldsymbol{\lambda}) = f_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}_i(\boldsymbol{z}_i, \boldsymbol{x}_i)$ may be regarded as a local subsystem Lagrangian function. Thus we can replace L by N independent problems.

$$L(i) \quad \max_{\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i} l_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i, \boldsymbol{\lambda}) \\ \text{s.t.} \quad \left.\begin{aligned} \boldsymbol{x}_i &= \boldsymbol{T}_i(\boldsymbol{z}_i, \boldsymbol{m}_i), \\ \boldsymbol{h}_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) &\leqslant \boldsymbol{0}. \end{aligned}\right\} \tag{7i}$$

The information flow between D and $L(i)$ at each iteration is shown in figure 3.

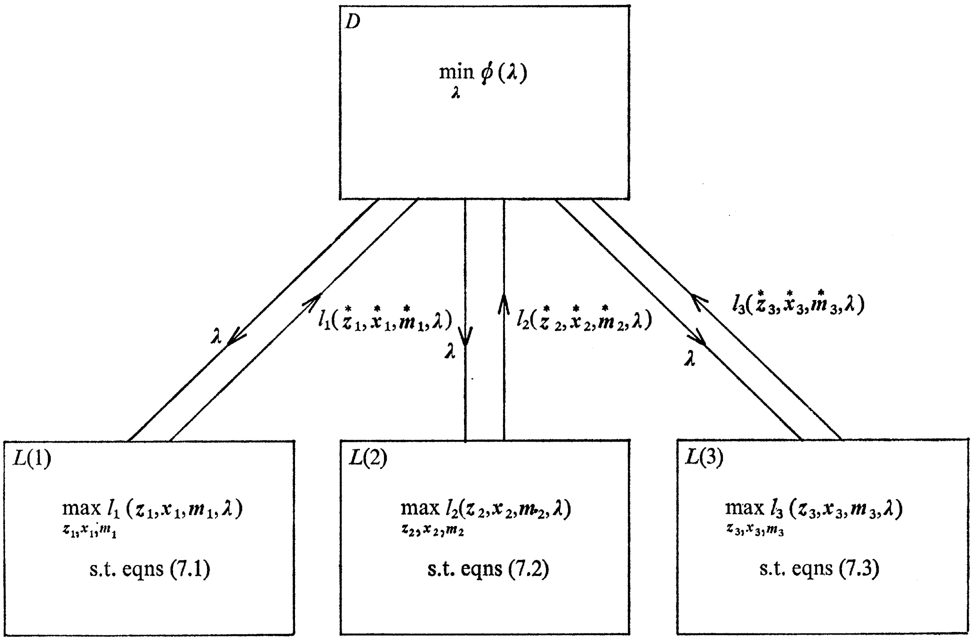
D
$\min_{\lambda} \phi(\lambda)$
λ
$l_1(\overset{*}{z}_1, \overset{*}{x}_1, \overset{*}{m}_1, \lambda)$
λ
$l_2(\overset{*}{z}_2, \overset{*}{x}_2, \overset{*}{m}_2, \lambda)$
$l_3(\overset{*}{z}_3, \overset{*}{x}_3, \overset{*}{m}_3, \lambda)$
λ
$L(1)$
$\max_{z_1, x_1, m_1} l_1(z_1, x_1, m_1, \lambda)$
s.t. eqns (7.1)
$L(2)$
$\max_{z_2, x_2, m_2} l_2(z_2, x_2, m_2, \lambda)$
s.t. eqns (7.2)
$L(3)$
$\max_{z_3, x_3, m_3} l_3(z_3, x_3, m_3, \lambda)$
s.t. eqns (7.3)

FIGURE 3. Information flow in dual coordination.

In the context of interconnected production systems introduced in §3*a*, the dual decomposition presented above has a direct economic interpretation (Brosilow & Lasdon 1965). Note that each element of the vector $G \begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{bmatrix}$ on the right and side of (3) is associated with just one product stream. If we assign a price to each stream and ask that each subsystem which makes or uses this stream shall sell it or buy it at this price, and if the profit or loss thereby incurred is added to or subtracted from the local objective (profit) function, we find that each local objective function so modified becomes

$$f_i(\boldsymbol{z}_i,\ \boldsymbol{x}_i,\ \boldsymbol{m}_i) - \boldsymbol{\lambda}^{\mathrm{T}} G_i \begin{bmatrix} \boldsymbol{z}_i \\ \boldsymbol{x}_i \end{bmatrix} = f_i(\boldsymbol{z}_i,\ \boldsymbol{x}_i,\ \boldsymbol{m}_i) - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}_i(\boldsymbol{z}_i,\ \boldsymbol{x}_i),$$

where $\boldsymbol{\lambda}$ is the vector of stream prices. If we allow each subsystem to choose its own values for its inputs, outputs and controls to maximize this modified profit function subject only to its local constraints, we find we have formulated the Lagrangian subproblems $L(i)$. In this interpretation the rôle of the coordinator D is to adjust the prices $\boldsymbol{\lambda}$ so that each subsystem finds it profitable to produce and consume just the right amount of the product streams to maintain the interconnections in balance. In the light of this interpretation dual coordination is often called 'price coordination', and because the subsystem goals $f_i(\boldsymbol{z}_i,\ \boldsymbol{x}_i,\ \boldsymbol{m}_i)$ are modified by the terms $\boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}_i(\boldsymbol{z}_i,\ \boldsymbol{x}_i)$ through which the coordinator exerts its influence, the description 'goal coordination' is also used.

Analytically, dual coordination appears rather more attractive than primal coordination; in particular, only mild assumptions of continuity of f_i, $\boldsymbol{g}_i$, $\boldsymbol{h}_i$ are needed to ensure that the dual function $\phi(\boldsymbol{\lambda})$ is convex, and if the solutions of $L(i)$ are unique the gradient of $\phi(\boldsymbol{\lambda})$ exists and is given by $-\sum_{i=1}^{N} \boldsymbol{g}_i(\overset{*}{\boldsymbol{z}}_i,\ \overset{*}{\boldsymbol{x}}_i)$. Consequently reasonably fast optimization algorithms can be used to solve D – indeed it is possible to generate second derivative information at the local level for use in solving D by second order optimization algorithms (Foord 1974). Because the solution of problem D can be so tractable, dual coordination has received considerable attention in the literature which has not always reflected the limitations of the method. The obvious failing is that it is an infeasible method in the sense that the solutions $\overset{*}{\boldsymbol{z}}$, $\overset{*}{\boldsymbol{x}}$, $\overset{*}{\boldsymbol{m}}$ derived from any other value of $\boldsymbol{\lambda}$ than $\mathring{\boldsymbol{\lambda}}$ do not satisfy the interconnection constraints (5), but this is a limitation only in some on-line situations. More importantly the presentation up to this point has not considered whether $\mathring{\boldsymbol{\lambda}}$ exists. If it does then solution of the dual problem D will find it because $\phi(\boldsymbol{\lambda})$ is convex, but in many problems it is observed that for the optimal solution $\overset{*}{\boldsymbol{\lambda}}$ of D, the corresponding solutions of the subproblems $L(i)$ are not unique and none of the values of $\overset{*}{\boldsymbol{z}}$, $\overset{*}{\boldsymbol{x}}$, $\overset{*}{\boldsymbol{m}}$ satisfy the constraints (5). In this situation the dual function is not differentiable, $\overset{*}{\boldsymbol{\lambda}}$ cannot be identified with $\mathring{\boldsymbol{\lambda}}$ and the optimal value of the dual function $L(\overset{*}{\boldsymbol{z}},\ \overset{*}{\boldsymbol{x}},\ \overset{*}{\boldsymbol{m}},\ \overset{*}{\boldsymbol{\lambda}})$ does not equal F. Sufficient conditions on the problem M to ensure that this kind of failure does not occur in dual coordination, are that the functions f_i are strictly concave, $\boldsymbol{h}_i$ are convex and $\boldsymbol{T}_i$ are linear. These conditions are restrictive for practical problems; they are however sufficient conditions only, and it is quite possible that problems not satisfying these conditions can be solved by using dual coordination. For example Javdan (1976) has shown that certain problems with quadratic objectives and quadratic equality constraints can be solved satisfactorily by using dual coordination, but it must be emphasized that in the absence of special circumstances the application of dual coordination to problems which do not satisfy the sufficient conditions given above cannot be robust and failure is highly probable (Foord 1974; Simmons 1975).

4. The relevance to industry of the basic methods of decentralized optimization

The broad conclusions to be drawn from the discussion given in §3 are that in a real industrial environment, coordination by specifying production targets or product transfer prices may be difficult to implement or may not lead to the optimal policy: the aforementioned assumptions of convexity and linearity which we required to make progress with solving the mathematical programming problem M introduced in §3*a* seem restrictive and therefore the theory would appear to offer little help in guiding the solution of real problems. While we believe these conclusions are largely valid in the context given, we also believe that they are too sweeping, and based on too narrow a concept of the problem, to warrant total rejection of the decentralized or hierarchical approach. Notwithstanding the difficulties of the theory given in §3, the observations made in §2 still stand – it is observed that industrial systems are indeed hierarchical and that decentralized optimization is made to work well enough though possibly not truly optimally. This last point is crucial; the theory of decentralized optimization as discussed in much of the literatiure and as presented in §3 is concerned with the optimal solution of a precisely defined problem, whereas in industry we require an adequate solution to a rather poorly or approximately defined problem. If this pragmatic view of industrial problems is accepted then we can allow ourselves greater flexibility both in formulating the mathematical programming problem and in devising viable decentralized solution techniques. We discuss first the central significance of problem formulation and, in particular, of the kind of models which may be used to represent the salient features of a system.

Faced with the very great difficulty of modelling a large complex system, but given that we are not expected to produce exact or optimal results, we should not build ever more complex models involving large numbers of parameters, perhaps with stochastic characteristics, which we have little chance of quantifying. Rather the procedure should be to build the simplest model that will suffice and that will suit the needs of any mathematical techniques we have to use on it. Since we only require an approximation to the optimal solution of our problem we are at liberty to make reasonable amendments to our problem formulation to suit our needs and overcome some of the difficulties in our solution techniques. For example, Drew (1975) in his approach to job shop scheduling by decentralized optimization has illustrated how the introduction of reasonable approximation led to great problem simplification without incurring serious loss of optimality. This process of 'problem evolution' has been discussed and illustrated by Wilson (1977).

In pursuing this approach to modelling it is natural to consider first linear or piecewise linear models. For example, it was mentioned in §2 that the long-term planning models for the petrochemical industry were linear, and linear models are widely and successfully used elsewhere in industry. If f_i, $\boldsymbol{T}_i$ and $\boldsymbol{h}_i$ are linear then M can be solved very adequately by linear programming, and if a decentralized approach is required the method of Dantzig & Wolfe (1960) and related methods (Lasdon 1970) are available. At the next stage of the problem evolution quadratic objective functions f_i may be introduced; the conditions for successful application of both primal and dual coordination are now satisfied, but it should be noted that, with f_i quadratic and $\boldsymbol{h}_i$ and $\boldsymbol{T}_i$ linear, M is a classical quadratic programming problem for which effective centralized solution techniques exist exploiting sparse matrix methods (Gill & Murray 1974). The motivation for decomposing such a problem will arise more from the demands of a hierarchical management structure as discussed in §2 than from any computational considerations. Linear and quadratic programming algorithms comprise by far the majority of mathematical programming techniques exploited in industry. Nonlinear programming problems are infrequent, but nevertheless there do arise from time to time large scale nonlinear problems which cannot be satisfactorily represented by linear approximations (for example problems containing quotients or products of important variables having a wide variation of magnitude) or wherein quadratic or even strictly convex objective functions are not permissible. In general these problems can be tackled only clumsily by linear or quadratic programming and yet it is just these problems which cause so much difficulty with the coordination methods of §3. However, provided one can move away from the rigid formulation of M given in equations (1), (2), (5) and (6), viable decomposition algorithms can be derived and two such algorithms are outlined in §5. They are both based upon the idea of relaxing in some way the equality coupling constraints (5), the justification for this relaxation coming from the considerations of realistic approximation and problem evolution discussed above. The first method, related to primal coordination, and of wide applicability, assumes that the interconnection constraints need not be satisfied exactly, but simply 'closely'; whereas the second method, based directly on dual coordination, explicitly recognizes the time element and the existence of storage and assumes that the interconnection constraints need only be satisfied on average. We use the adjective 'slack' to describe constraints which may be satisfied in either of these ways.

5. Coordination of systems with slack interconnections

(*a*) *A penalty function method of coordination*

Findeisen (1974*b*) aptly describes the penalty function method as a means of avoiding many of the difficulties of primal coordination by introducing the interconnection constraints 'softly': this soft introduction of the interconnection constraints is achieved by adjoining to the global objective function a term which penalizes dissatisfaction of these constraints. Exact satisfaction of these constraints is then no longer demanded, but close satisfaction can be enforced by making the penalty term large enough. In this way the problem associated with inadequate degrees of freedom of the local problems $P(i)$, and with the determination of the feasible set V, are largely ameliorated. Coordination by penalty function methods has been studied in detail by Tatjewski (1974), and has been considered for the optimization of dynamic systems by Pearson (1971) who calls the method pseudo-model coordination. There is some flexibility in the way in which the penalty term is constructed, and this allows some variation in formulating the coordination procedure: unfortunately the literature does not always distinguish between these variations and this has led to some confusion about the properties of the method. The following treatment is due to Simmons (1976).

Again for clarity of illustration, attention is concentrated on the production system introduced in §3*a*, but the representation of the interconnection constraints by equation (5) is inconvenient, and a different partitioning of the basic constraint equation (3) is used here. Because each input is uniquely associated with an interconnection stream, each element of $\boldsymbol{z}$ occurs only once in the equations (3), each element in a separate equation; therefore these equations can be solved for $\boldsymbol{z}$ to give

$$\boldsymbol{z} = \boldsymbol{C}\boldsymbol{x}, \tag{14}$$

where C is a matrix of ones and zeros. By suitably partitioning C into N groups of columns and N groups of rows equation (14) can be written

$$\begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \\ \boldsymbol{z}_N \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \dots C_{1N} \\ C_{21} & C_{22} \dots C_{2N} \\ \vdots & \vdots \qquad \vdots \\ C_{N1} & C_{N2} \dots C_{NN} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix}$$

which can be rewritten

$$\boldsymbol{z}_i = \sum_{j=1}^{N} C_{ij}\, \boldsymbol{x}_j, \tag{15}$$

where the C_{ij} are the submatrices of C conformable with $\boldsymbol{z}_i$ and $\boldsymbol{x}_j$.

By using this form of the interconnection constraint equation, the global problem becomes M'

$$M' \quad \max_{\boldsymbol{z},\boldsymbol{x},\boldsymbol{m}} \sum_{i=1}^{N} f_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i)$$

$$\left.\begin{aligned} \text{s.t.} \quad \boldsymbol{z}_i &= \sum_{j=1}^{N} C_{ij}\, \boldsymbol{x}_j \\ \boldsymbol{x}_i &= \boldsymbol{T}_i(\boldsymbol{z}_i, \boldsymbol{m}_i) \\ \boldsymbol{h}_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) &\leqslant \boldsymbol{0} \end{aligned}\right\} i = 1, 2, \dots N.$$

As before we seek a way of decomposing this problem into N or more interacting but simpler subproblems. The first step is to introduce auxiliary variables $\boldsymbol{u}$ which are required at this stage to be exactly equal to the outputs $\boldsymbol{x}$. Problem M' can now be transformed quite trivially into the exactly equivalent problem M''

$$M'' \quad \max_{\boldsymbol{z},\boldsymbol{x},\boldsymbol{m},\boldsymbol{u}} \sum_{i=1}^{N} f_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i)$$

$$\left.\begin{aligned} \text{s.t.} \quad \boldsymbol{z}_i &= \sum_{j=1}^{N} C_{ij}\, \boldsymbol{u}_j \\ \boldsymbol{x}_i &= \boldsymbol{T}_i(\boldsymbol{z}_i, \boldsymbol{m}_i) \\ \boldsymbol{h}_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) &\leqslant \boldsymbol{0} \\ \boldsymbol{x}_i &= \boldsymbol{u}_i \end{aligned}\right\} i = 1, 2, \dots N.$$

In the context of the production subsystem shown in figure 1 this reformulation implies that each output stream $\boldsymbol{x}_i$ is cut, and the stream value beyond the cut is denoted by $\boldsymbol{u}_i$ as shown in figure 4.

This conceptual cut is of no significance as long as the constraint $\boldsymbol{x}_i = \boldsymbol{u}_i$ is satisfied; however, in the next stage of problem evolution this constraint is dropped but its close satisfaction is ensured by adding to the objective a penalty function which is quadratic in the constraint dissatisfaction. Thus consider the problem $M(K)$ which is derived by using an exterior penalty function technique to deal with the constraint $\boldsymbol{x} = \boldsymbol{u}$.

$$M(K) \quad \max_{\boldsymbol{z},\boldsymbol{x},\boldsymbol{m},\boldsymbol{u}} \sum_{i=1}^{N} f_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) - \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{u}_i)^{\mathrm{T}} K_i (\boldsymbol{x}_i - \boldsymbol{u}_i),$$

$$\left.\begin{aligned} \text{s.t.} \quad \boldsymbol{z}_i &= \sum_{j=1}^{N} C_{ij}\, \boldsymbol{u}_j \\ \boldsymbol{x}_i &= \boldsymbol{T}_i(\boldsymbol{z}_i, \boldsymbol{m}_i) \\ \boldsymbol{h}_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) &\leqslant \boldsymbol{0} \end{aligned}\right\} i = 1, 2, \dots, N,$$

where the K_i are diagonal matrices of positive constants. The solution of the problem $M(K)$ will not be the same as that of M'' (and M') but it can be made close by making the matrices K_i large, and as $K_i \to \infty$ the solutions of the two problems and the optimal objective function values become identical (Fiacco & McCormick 1968). For very large values of K_i the problem $M(K)$ is ill conditioned and its numerical solution is likely to be slow from arbitrary starting points. Thus the usual treatment of constraints by this exterior penalty function method is to solve a sequence of problems $M(K)$ with monotonically increasing values of K_i, starting each problem from the solution point of the previous problem in the sequence. However, if the arguments given in §4 are valid and some degree of relaxation of the interconnection constraints is permitted, it will not be necessary to pursue a sequence of solutions as $K_i \to \infty$; the solution of the problem $M(K)$ for a fixed moderate value of the K_i will be acceptable.

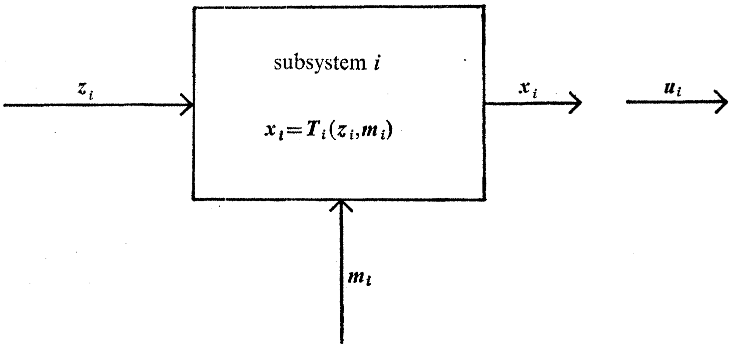

FIGURE 4. Subsystem with cut output stream.

The required solution of $M(K)$ can be found in a decentralized fashion by the parametric decomposition or primal coordination methods of §3*b*. For a fixed value of $\boldsymbol{u}$ the problem $M(K)$ separates into the N subproblems $Q(i)$

$$Q(i) \quad \max_{\boldsymbol{z}_i,\, \boldsymbol{x}_i,\, \boldsymbol{m}_i} f_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) - (\boldsymbol{x}_i - \boldsymbol{u}_i)^{\mathrm{T}} K_i (\boldsymbol{x}_i - \boldsymbol{u}_i)$$

$$\left.\begin{aligned} \text{s.t.} \quad \boldsymbol{z}_i &= \sum_{j=1}^{N} C_{ij}\, \boldsymbol{u}_j \\ \boldsymbol{x}_i &= \boldsymbol{T}_i(\boldsymbol{z}_i, \boldsymbol{m}_i) \\ \boldsymbol{h}_i(\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{m}_i) &\leqslant \boldsymbol{0} \end{aligned}\right\}. \tag{16i}$$

Let the solution of these N subproblems be denoted by $\overset{*}{\boldsymbol{z}}_i$, $\overset{*}{\boldsymbol{x}}_i$, $\overset{*}{\boldsymbol{m}}_i$, all of which will be functions of the fixed value of $\boldsymbol{u}$. The value of the global objective function of $M(K)$ for these values of inputs, outputs and controls is defined as $\theta(\boldsymbol{u})$ where

$$\theta(\boldsymbol{u}) = \sum_{i=1}^{N} \{f_i(\overset{*}{\boldsymbol{z}}_i, \overset{*}{\boldsymbol{x}}_i, \overset{*}{\boldsymbol{m}}_i) - (\overset{*}{\boldsymbol{x}}_i - \boldsymbol{u}_i)^{\mathrm{T}} K_i (\overset{*}{\boldsymbol{x}}_i - \boldsymbol{u}_i)\}.$$

The optimal value of $\boldsymbol{u}$ is then determined by solution of the problem

$$U \quad \max_{\boldsymbol{u}} \theta(\boldsymbol{u})$$

$$\text{s.t. } \boldsymbol{u} \in W,$$

where the set W is defined

$$W = \{\boldsymbol{u} | \text{there exist } \boldsymbol{z},\ \boldsymbol{x}, \text{ and } \boldsymbol{m} \text{ satisfying (16i) for all } i\}.$$

Note that this set W contains the set X of all feasible values of $\boldsymbol{x}$ as a subset, and without serious loss of freedom W may be replaced by X: in contrast to the set V introduced for primal coordination in §3*b*, the set X is easily determined. Equally in comparing problems $Q(i)$ with problems $P(i)$ we note that the introduction of the auxiliary coordinating variables $\boldsymbol{u}$ has given more degrees of freedom to the problems $Q(i)$ in that both $\boldsymbol{m}_i$ *and* $\boldsymbol{x}_i$ may be freely chosen in $Q(i)$ since only $\boldsymbol{z}_i$ is uniquely fixed by the coordinator through equation (15).

Given that the functions f_i and $\boldsymbol{T}_i$ are differentiable and the functions $\boldsymbol{h}_i$ are continuous and functions of $\boldsymbol{m}_i$ only and given that the solutions to problems $Q(i)$ are unique it can be shown that the derivatives of $\theta(\boldsymbol{u})$ exist, and therefore gradient methods of optimization may be used to solve U. Indeed, except at points where the active set of the constraints $\boldsymbol{h}_i$ are changing, the second derivatives of $\theta(\boldsymbol{u})$ exist and may be used in second order methods of solving U. These considerations show that the penalty function method of coordination is of wider applicability than the basic primal coordination method and can be computationally much more efficient, although numerical experiments show that the method is still no quicker than a centralized solution of the global problem and therefore the decision to adopt the penalty function method can still only be motivated by the original hierarchical environment of the problem.

(*b*) *The randomized solution method*

It was stated in §3*c* that in the absence of certain rather restrictive convexity and linearity assumptions, dual coordination could fail. However, the concept of 'randomized solutions', originally due to Whittle (1971), shows that dual coordination need not be abandoned when failure occurs provided that it is necessary only to satisfy the equality constraints (5) on average. In the context of the production system of §3*a*, Simmons (1975) has suggested that the averaging can be achieved by the provision of storage for every stream; the need for instantaneous satisfaction of the constraints is thus relaxed but if the constraints are satisfied on average it is possible to ensure that the storage will be neither saturated nor depleted over a period of time.

The theory of the method is based quite simply on the observation made in §3*c* that dual coordination only fails when for $\boldsymbol{\lambda} = \overset{*}{\boldsymbol{\lambda}}$, where $\overset{*}{\boldsymbol{\lambda}}$ solves D, the solution of the problems $L(i)$, and therefore the solution of the problem L, is not unique. Let the multiple solutions to L be denoted by $\overset{*}{\boldsymbol{z}}{}^k$, $\overset{*}{\boldsymbol{x}}{}^k$, $\overset{*}{\boldsymbol{m}}{}^k$ and let

$$\sum_{i=1}^{N} \boldsymbol{g}_i(\overset{*}{\boldsymbol{z}}{}_i^k, \overset{*}{\boldsymbol{x}}{}_i^k) = \boldsymbol{b}^k, \tag{17}$$

where $k = 1, 2, 3, \ldots$

Each vector $\boldsymbol{b}^k$ may be interpreted as the constraint dissatisfaction vector corresponding to each solution of L. Whittle and Simmons show that it is always possible to find a convex combination of some or all of these vectors $\boldsymbol{b}^k$ which equals the null vector, that is to say it is possible to find numbers s_j such that

$$\sum_{j=1}^{r} s_j\, \boldsymbol{b}^j = \boldsymbol{0}, \tag{18}$$

$$\sum_{j=1}^{r} s_j = 1, \quad s_j \geqslant 0 \quad (j = 1, 2, \ldots, r), \tag{19}$$

where $2 \leqslant r \leqslant c+1$ and where c is the number of interconnection streams, i.e. the dimension of the vectors $\boldsymbol{b}^j$. The immediate interpretation of these results is that if the subsystems are operated in a random or cyclic sequence of r states in such a way that they spend a time proportional to s_1 in the state determined by $\overset{*}{\boldsymbol{z}}{}^1$, $\overset{*}{\boldsymbol{x}}{}^1$, $\overset{*}{\boldsymbol{m}}{}^1$, a time proportional to s_2 in the state determined by $\overset{*}{\boldsymbol{z}}{}^2$, $\overset{*}{\boldsymbol{x}}{}^2$, $\overset{*}{\boldsymbol{m}}{}^2$ and so on over the r states, then equations (18) and (19) imply that the constraint dissatisfaction arising in these r states averages to zero. Moreover, the theory shows that the average of the objective function values achieved in each of the r states is equal to the dual function, namely

$$\sum_{j=1}^{r} s_j \left\{ \sum_{i=1}^{N} f_i(\overset{*}{\boldsymbol{z}}{}_i^j, \overset{*}{\boldsymbol{x}}{}_i^j, \overset{*}{\boldsymbol{m}}{}_i^j) \right\} = \phi(\overset{*}{\boldsymbol{\lambda}}) \tag{20}$$

and hence it follows from the inequality (12) that

$$\sum_{j=1}^{r} s_j \left\{ \sum_{i=1}^{N} f_i(\overset{*}{\boldsymbol{z}}{}_i^j, \overset{*}{\boldsymbol{x}}{}_i^j, \overset{*}{\boldsymbol{m}}{}_i^j) \right\} \geqslant F. \tag{21}$$

Thus the average value of the objective function cannot be less than, and will usually exceed, the maximum value F which can be achieved when the constraints (5) are instantaneously satisfied all the time.

Before this useful extension of dual coordination can be exploited it is necessary to find a satisfactory numerical method of solving the dual problem D, because as stated in §3*c*, when there are multiple solutions to the problem L, the dual function is not differentiable. However, as $\phi(\boldsymbol{\lambda})$ is convex under the mild assumptions of continuity of the functions f_i, $\boldsymbol{h}_i$ and $\boldsymbol{g}_i$, and as the vectors $\boldsymbol{b}^j$ introduced above are subdifferentials of ϕ and are easily computed, the convex programming method being developed by Lemarechal (1975) and Wolfe (1975) may be useful – indeed the method of these authors might also offer some hope of solving efficiently the problems *CP* and *U* under more general assumptions that have been possible hitherto, but the method is not yet fully tested. Simmons has preferred to solve D by Kelley's cutting plane method (Kelley 1960; Marsten 1975) which approximates the dual function by an increasing set of supporting hyperplanes. At each step of the algorithm the lowest vertex of the polyhedron bounded by the planes is found by linear programming, and a new plane supporting the dual function at this solution point is added. As the algorithm progresses the sequence of solution points found by linear programming at each step converges uniformly on the solution $\phi(\overset{*}{\boldsymbol{\lambda}})$. The method has only linear convergence and is slow, but by solving the dual of the linear program at each step, values of s_i and b^j satisfying (18) and (19) are obtained directly, even though the $\boldsymbol{b}^j$ correspond to a value of $\boldsymbol{\lambda}$ which is not the optimal value $\overset{*}{\boldsymbol{\lambda}}$. Thus at least as the algorithm progresses it gives an operating strategy which converges steadily to the optimal, and which ensures that the interconnection constraint violations average to zero at every step of the algorithm; thus on-line use of the method is acceptable.

6. Conclusions

The methods of §5*a* and *b* overcome some of the objections to the basic methods of primal and dual coordination given in §3*b* and *c*, and yet they retain the fundamental features of the basic methods in that the penalty function method coordinates by specifying output targets (the $\boldsymbol{u}_i$) and the randomized solution method coordinates by specifying prices. What remains to be assessed is whether the assumptions of constraint slackness used in these methods are valid in practice. Assessment of the numerical robustness and indeed of the whole feasibility and value of these methods can only be made by applying them to real industrial problems. Unfortunately, opportunities for such practical studies are rare and there are many reasons for this. It must be accepted that the remarks made in §4 are true – there are few occasions in industry when large-scale nonlinear programming problems need to be solved, but on the occasions when such problems do arise, human inertia and resistance to change make it unlikely that these methods will be tried, or even considered. It is noteworthy that even the well established method of Dantzig & Wolfe (1960) for decomposing linear programs is little used, even though suitable large scale linear programs occur widely in industry. In reviewing just this situation, Orchard–Hays (1973) puts forward as one of the main reasons for the lack of exploitation of decomposition, the power and effectiveness of present day, single level, linear programming packages, and the speed and size of the computers available to run these packages: by using these facilities, industry can solve its mathematical programming problems more or less adequately even though on occasions, somewhat inefficiently and inaccurately.

To be fair, it must be emphasized that the promising decentralized optimization techniques, which have formed the main topic of this paper, tackle only one aspect of exploiting structure in industrial problems. The important and largely unsolved problems of decentralized information structures has already been mentioned in §1. Equally important is the problem of determining structure – a problem which was glossed over in §3*a* where it was assumed that the delineation of a subsystem and identification of its inputs and outputs was trivially obvious: in practical problems, particularly where the interconnections are not readily recognizable streams, such as in socio-economic systems, the analysis of structure to reveal not only the 'natural' subsystems and their interconnections but also the strength or significance of these interconnections is an essential step and techniques for carrying out this analysis are being developed (Kevorkian & Snoek 1973; Gourlay, McLean & Shepherd 1977). By such analysis techniques it may be possible to devise decentralized optimization techniques which coordinate with respect to the 'significant interconnections' only and this is an interesting line of further research (Findeisen 1975*b*).

Thus in the subject area of complex systems there is already a considerable body of theory and there is a substantial program of continuing research. At this stage it is essential to obtain from industry some definite feedback on the value and effectiveness of the ideas and techniques being developed, otherwise the whole subject area will become yet another branch of pure mathematics.

References (White & Simmons)

Athans, M. 1974 Survey of decentralized control methods. *3rd NBER/FRB Workshop on Stochastic Control Washington D.C.*

Bailey, F. N. 1976 Decision processes in organisations. In *Large-scale dynamical systems* (ed. R. Saeks). California: Point Lobos Press.

Benders, J. F. 1962 Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik* **4**, 238–252.

Brosilow, C. B. & Lasdon, L. S. 1965 A two level optimization technique for recycle processes. *Proceedings of the A.I.Ch.E. – Ind. Chem. Eng. Joint Meeting, London* **4**, 75–**4**, 83.

Cheliustkin, A. & Lefkowitz, I. 1975 State of the art review of integrated systems control in the steel industry. *Report WP*-75-62. Schloss Laxenburg: I.I.A.S.A.

Dantzig, G. B. & Wolfe, P. 1960 Decomposition principle for linear programs. *Operations Res.* **8**, 101–111.

Drew, S. A. W. 1975 A study in the application of large scale control methods to a practical industrial problem. *Proceedings of the 6th IFAC World Congress, Boston* **3.1**, 1–**3.1**, 9.

Dyer, P. 1976 Application of mathematical models in a petrochemical business. *Proceedings A.I.Ch.E. Meeting, Atlantic City*.

Fiacco, A. V. & McCormick, G. P. 1968 *Nonlinear programming: sequential unconstrained minimization techniques*. New York: Wiley.

Findeisen, W. 1974*a* *Wielopoziomowe układy sterowania*. Warsaw: Panstwowe Wydawnictwo Naukowe.

Findeisen, W. 1974*b* Control and coordination in multilevel systems. *2nd Polish–Italian Conference on Applications of Systems Theory to Economy, Management and Technology, Pugnoschiuso*; see also *Technical Report* no. 5/1974, Institute of Automatic Control, Technical University of Warsaw.

Findeisen, W. & Lefkowitz, I. 1969 Design and application of multilayer control. *Proceedings of the 4th IFAC World Congress, Warsaw.*

Foord, A. G. 1974 On-line optimisation of a petrochemical complex. Ph.D. thesis, Cambridge University.

Geoffrion, A. M. 1970 Elements of large-scale mathematical programming. *Management Sci. Theory* **16**, 652–691.

Gill, P. E. & Murray, W. (eds) 1974 *Numerical methods for constrained optimization.* London: Academic Press.

Gourlay, A. R., McLean, M. & Shepherd, P. 1977 The identification and analysis of sub-system structure of models. *Appl. Math. Modelling* **1**, 245–252.

Himmelblau, D. M. (ed.) 1973 *Decomposition of large-scale problems.* Amsterdam: North-Holland.

Ho, Y. C. & Chu, K. C. 1974 Information structure in dynamic multi-person control problems. *Automatica* **10**, 341–351.

Javdan, M. R. 1976 Extension of dual coordination to a class of non-linear systems. *Int. J. Control* **24**, 551–572.

Kelley, J. E. 1960 The cutting-plane method for solving convex programs. *J. Soc. Ind. Appl. Math.* **8**, 703–712.

Kevorkian, A. K. & Snoek, J. 1973 Decomposition in large-scale systems theory and applications. In *Decomposiion of large-scale problems* (ed. D. M. Himmelblau). Amsterdam: North Holland.

Kulikowski, R., Krus, L., Manczak, K. & Straszak, A. 1975 Optimization and control problems in large-scale systems. *Proceedings of the 6th IFAC World Congress, Boston* **19.1**, 1–**19.1**, 13.

Lasdon, L. S. 1970 *Optimization theory for large systems.* London: Macmillan.

Lefkowitz, I. 1966 Multilevel approach applied to control system design. *Trans. A.S.M.E.* **88**, 392–398.

Lemarechal, C. 1975 An extension of Davidon methods to non-differentiable problems. *Mathematical Programming Study* no. 3, 95–109.

Marsten, R. E. 1975 The use of the Boxstep method in discrete optimization. *Mathematical Programming Study* no. 3, 127–144.

Mesarovic, M. D., Macko, D. & Takaraha, Y. 1970 *Theory of hierarchical multilevel systems.* New York: Academic Press.

Orchard-Hayes, W. 1973 Practical problems in LP decomposition. In *Decomposition of large-scale problems* (ed. D. M. Himmelblau). Amsterdam: North Holland.

Pearson, J. D. 1971 Dynamic decomposition techniques. In *Optimization methods for large-scale systems* (ed. D. A. Wismer). New York: McGraw Hill.

Rosen, J. E. 1964 Primal partition programming for block diagonal matrices. *Numerische Mathematik* **6**, 250–260.

Sandell, N. R., Varaiya, P. P. & Athans, M. 1976 A survey of decentralised control methods for large-scale systems. *Proceedings IFAC Symposium on large-scale Systems Theory and Applications, Udine.*

Simmons, M. D. 1975 The decentralised profit maximisation of interconnected production systems. *Report No. CUED/F – Control/TR*101. Cambridge University Engineering Department.

Simmons, M. D. 1976 Comments on 'A new algorithm...'. *Int. J. Control*, **24**, 441–442.

Steel Industry Project Staff 1975 Hierarchical control systems in the steel industry. *Proceedings of the 6th IFAC World Congress, Boston* **12.1**, 1–**12.1**, 14.

Tatjewski, P. 1974 A penalty method in multilevel optimisation and its applicability conditions. *Technical Rep.* 11/74. Technical University of Warsaw: Institute of Automatic Control.

Whittle, P. 1971 *Optimization under constraints.* London: Wiley.

Wilson, I. D. 1977 *Decentralised control based on problem decomposition.* Ph.D. thesis, Cambridge University.

Wolfe, P. 1975 A method of conjugate subgradients for minimising nondifferentiable functions. *Mathematical Programming Study* no. 3, 145–173.