SOME

CONCEPTUAL FOUNDATIONS OF SYSTEMS AND COMPUTER SCIENCE

by

Sanjoy K. Mitter

Department of Electrical Engineering and Computer Science
and
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, Mass. 02139
U.S.A.

I. Introduction

The dominant paradigm for much of Science has been the paradigm of Physics, namely, the discovery of the laws of nature and their encapsulation in the form of mathematical equations, which are local in character. The laws of physics are immutable and the discovery of the laws of nature proceeds from theories (usually mathematical) which are then verified by experiments which are, in principle, repeatable infinitely often under identical conditions. So dominant is the paradigm of physics, that even a field such as psychoanalysis is under intense scrutiny today with the methods and standards of physics.(1) The fact that this may be inappropriate and that an observational science such as astronomy may be a better model for psychoanalysis does not seem to be taken into consideration. In a similar vein, much of the difficulties of the conceptual foundation of economics could be traced to the chains put on the subject of having Physics as its primary paradigm.

The "unusual success of mathematics"(2) in Physics is so pervasive that one cannot conceive of the science of physics without mathematics. This "mathematization of nature",(3) which began at least with Galileo, has continued for centuries, and after a period of temporary decline in the fifties and sixties is again in full bloom.(4)

Physics as we have said is inseparable from Mathematics. Conversely, problems of physics have given rise to new mathematics. Geometry, ordinary and partial differential equations, functional analysis (especially the theory of unbounded operators on a Hilbert space, operator algebras), theory of group representations, all have deep roots in physics. In a very profound sense, physics, its mathematization and the interpretation of nature are all different facets of the same fundamental reality.

Much of physics deals with systems which are isolated and the question of studying the effects of external influences on the dynamical system is not of concern. Even when external influences are allowed, they are typically taken to be functions of the state of the system and hence can be incorporated into the original dynamical description of the system. The prime example of this situation is a conservative mechanical system where external influences are derived from a potential function. Thus, although Newton's second law:

$$m\frac{d^2q}{dt^2} = F \quad , \tag{1.1}$$

where m is the mass, q the configuration variable and F the external force, was originally written with the external influence, the force, in a predominant position, the modern study of classical mechanics almost exclusively concentrates on the Hamiltonian formalism:

$$\left\{ \begin{aligned} \frac{dq}{dt} &= \frac{\partial H\cdot(q,p)}{\partial p} \\ \frac{dp}{dt} &= -\frac{\partial H\cdot(q,p)}{\partial p} \end{aligned} \right. \tag{1.2}$$

where p is the conjugate variable and H is the Hamiltonian of the system. A typical Hamiltonian is of the form

$$H(q,p) = V(q) + p^2 \quad ,$$

where V is a potential function, and it is assumed that the Hamiltonian is a conserved quantity along the solutions of Hamilton's equations (1.2). This viewpoint is embodied in the mathematical formalization of classical mechanics as symplectic geometry.(5)

Secondly, much of physics is concerned with equilibrium or quasi-state situations. This is the so-called thermodynamic formalism.(6) To obtain this equilibrium description, one has to often pass to the thermodynamic (infinite-volume) limit as is customarily done in equilibrium statistical mechanics. However, in spite of many efforts, the conceptual foundations of classical thermodynamics, notably the derivation of the thermodynamic laws from dynamical descriptions, has not been satisfactorily achieved in the physics literature.

Thirdly, the idea of making inaccurate dynamical measurements on functions of the "state" of the system and making "inferences" about the state of the system from these inaccurate observations, is notably absent from physics.

As we have remarked, the fundamental goal of physics is to understand the laws of nature. It thus basically is concerned with analysis and the process of this analysis is one of discovery. There is no doubt, that the fundamental laws of nature do exist, they always have existed, and they will continue to exist in all future. There is however a need for a different paradigm when one is interested in systems which are not of nature but man-made, where one might want to create a new device to perform a specific function, or shape the national economy to grow along a particular path or to synthesize a complex system consisting of interconnections of subsystems to perform a complex task. These systems are not isolated in the sense that they interact with an external environment, they do have inputs or external influences, some of which can be controlled and some of which are uncontrolled, the behavior of the system can be observed (perhaps inaccurately) and finally the behavior of the system can be changed by a feedback mechanism which feeds the input via a control mechanism into the system. The concern here is one of synthesis (not of analysis) and the process is one of invention and not of discovery.

For a mathematical theory of these systems and their synthesis, it should be obvious from the above discussion that we need to describe systems (physical or man-made) in terms of the external behavior of the system. We shall see that this viewpoint is illuminating even for systems which are usually of concern in physics. The internal description of systems is then inferred from the external description and leads to the notion of the "state" of a system and intuitively should contain the complete memory of the system. Once this is accomplished, the primary concerns of a theoretical framework are to understand the fundamental limitations of the system and to classify the system by describing a complete set of invariants (under appropriate transformations groups).

2. Fundamental Limitations of Systems

We want to illustrate the notion of fundamental limitation of systems by means of two examples. The first comes from communication theory and the second from linear systems theory. In discussing the second example, we shall also introduce the ideas of inputs, outputs and state of a system.

2.1 Heisenberg's Inequality, Band and Time Limited Functions

We start with the original description of Heisenberg's inequality in the context of one-dimensional quantum-mechanical particle.

A "state" of such a particle is a wave function $\psi \in L^2(\mathbb{R}')$ = (the space of square-integrable functions). The probability of finding the particle in the interval $a \leq x \leq b$ is

$$\int_a^b \psi^* \psi \, dx = \int_a^b |\psi|^2 \, dx \quad ,$$

where * denotes complex conjugate. The total probability is $\int_{\mathbb{R}} |\psi(x)|^2 \, dx$ which has to be 1.

An "observable" is a symmetric operator A acting on a suitable domain $D(A) \subset L^2(\mathbb{R}')$. The expectation value of A in the state ψ is defined to be

$$E(A) = \int \psi^* A \psi dx \qquad \text{for } \psi \in D(A) \quad .$$

The position of the particle corresponds to the operator multiplication by x. The "momentum" of the particle is associated with the operator $B\psi = (2\pi i)^{-1}\psi'$ acting on the domain, $D(B) = \{\psi \in L^2 | \int |\psi'| \, dx < \infty\}$ and where ' denotes differentiation. The expectation value of the power of the momentum operator in the state ψ is given by

$$\int \psi^* B^n \psi \, dx = \int \gamma^n |\hat{\psi}|^2 \, d\gamma$$

where $\hat{\psi}$ is the Fourier transform of ψ . Hence

$$\int_a^b \hat{\psi} \ \hat{\psi} \ d\gamma = \int_a^b |\hat{\psi}|^2 \, d\gamma$$

is the probability that the momentum finds itself in the interval $a \leq \gamma \leq b$.

The position and momentum satisfy the commutation relation

$$AB - BA = \frac{i}{2\pi}$$

and one can show

$$D[A - E(A)]^2 \times E[B - E(B)]^2 \geq \frac{1}{16\pi^2}$$

which has the interpretation that the position and the momentum cannot be measured simultaneously with arbitrary precision (the Uncertainty Principle). This principle is in fact an inequality involving Fourier transforms:

$$\left(\int x^2 |f(x)|^2 dx\right)\left(\int \gamma^2 |\hat{f}(\gamma)|^2 d\gamma\right) > \frac{1}{16\pi^2}\left(\int |f(x)|^2 dx\right)^2$$

A problem of interest in communication theory is that of <u>synthesizing</u> a signal f(t) with total power $\int |f(t)|^2 dt = 1$ with both

$$\alpha^2 = \int_{-a}^{a} |f(t)|^2 dt$$

and

$$\beta^2 = \int_{-b}^{b} |f(\gamma)|^2 d\gamma$$

as close to 1 as possible for fixed positive numbers a and b. $\alpha = 1$ means the signal is time-limited in the period $|t| \leqq a$ and $\beta = 1$ means the "power spectrum" of the signal is confined to the band $|\gamma| \leq b$. One can prove that it is impossible to make $\alpha = \beta = 1$.

A refined result says: the pairs α, β corresponding to actual signals f of unit power fill up the subregion of the unit square

$[0,1] \times [0,1]$ defined by

$$\cos^{-1}\alpha + \cos^{-1}\beta \geq \cos^{-1}\sqrt{\gamma_1}$$

(if α or $\beta = 0$ (resp. $= 1$) then the other is < 1 (resp. 0)), where $\gamma_1 = \sup[\alpha^2$ class of band-limited signals of unit power]. It is a function of the product ab only.

2.1 Invariants of Linear Systems(8)

We think of an input-output linear system as a linear causal map between a set of variables (time functions) called inputs (causes) and a set of variables (time functions) called outputs (effects). A mathematical description of such a linear system is given by

$$y(t) = \int_0^t W(t-\tau)\, u(\tau)\, d\tau \quad ,$$

where $y(t) \in \mathbb{R}^p$, $u(t) \in \mathbb{R}^m$ and $W(t-\tau) \in \mathscr{L}(\mathbb{R}^m, \mathbb{R}^p)$ ($p \times m$ matrix). The fact that the weighting function (kernel) depends on the difference $t - \tau$ reflects the fact that the system is time-invariant (shift-invariant). If we assume that the Laplace Transform of W is rational, then one can show that there exists a vector x of minimal dimension (say n), the state vector, and matrices A (nxn), B (nxm) and C (pxn) such that

$$\begin{cases} \dfrac{dx}{dt} = Ax(t) + Bu(t) \\ y(t) = Cx(t) \quad . \end{cases}$$

The initial condition is taken to be 0 which corresponds to the assumption that the system is initially at rest. Moreover, this representation is unique upto an isomorphism, in the sense that two such minimal systems are related by a similarity transformation on the space where x lives (i.e., $\mathbb{R}^n$).

In line with our discussion, we can ask the following question:

To what extent can we change the system by means of

(i) Coordinate transformations on the state space $\mathbb{R}^n$

(ii) Coordinate transformations on the input space $\mathbb{R}^n$

(iii) Feedback control of the form $u(t) = v(t) + Kx(t)$?

The mathematical question we are asking is the following:

Let M_n denote the set of nxn matrices and $M_{n,m}$ denote the set of nxm matrices and $M_{m,n}$ denote the set of mxn matrices.

Consider the group actions:

(a) $GL(n) \times M_n \to M_n$

: $(T,A) \mapsto T^{-1} AT$ where GL(n) is the group of nxn matrices which are invertible

(b) $GL(m) \times M_{n,m} \to M_{n,m}$

: $(S,B) \mapsto BS$, where GL(m) is the group of mxn matrices which are invertible

(c) $F \times (M_n) \to M_n$

: $(K,A) \mapsto A + BK$, where F is called the feedback group.

The simultaneous action of all these transformations is the semi-direct product of GL(n), GL(m) and F and is given by

$$(A,B) \longmapsto (T^{-1}(A + BK)T,\ T^{-1}BS) \quad .$$

If we denote by G this group, then the problem is to study the action of G on the space $\mathbb{R}^{n^2+mn}$ and classify the corresponding orbit space. One can show that this problem corresponds to the study of an algebraic vector bundle over $\mathbb{P}^1$, the complex projective space. According to a theorem of Grothendieck, every algebraic vector bundle over $\mathbb{P}^1$ is isomorphic to the direct sum of line bundles (i.e., vector bundles with one-dimensional fibers) and upto isomorphism, classes of algebraic vector bundles over $\mathbb{P}^1$ are in one-one correspondence with the set of integers $K, \leq \ldots \leq K_m$, $K_i \in Z$. In our problem, the integers K_i correspond to dimensions of certain subspaces of $\mathbb{R}^n$ and have the property $K_1 + K_2 + \ldots + K_m = n$. These are the invariants of the system and have the interpretation that these numbers cannot be changed by means of feedback control $u(t) = v(t) + Kx(t)$ in a coordinate free description of the linear system. It is another example of a fundamental limitation of a system.

3. Systems With External Variables(9)

What kind of procedure should we follow in trying to describe a system? The first step we have to take is to look at the system as an entity distinguished from the outside world. We have to make clear what belongs to the system and what we do not want to include in it. After this separation between system and environment has been accomplished, we have, roughly speaking, the following three possibilities to describe the system.

The first one is that we consider the system as actually isolated from the outside world, or at least that for all purposes of accuracy we may regard the system as isolated. The paradigmatic example of this possibility is our solar system. Indeed this can be regarded as a world on its own. However, it is hard to find down-to-earth and real (i.e., not idealized) systems which have this same strictly isolated behavior although it may be in many instances a reasonable assumption.

A second possibility is to regard the part of the outside world which may influence the system under consideration as nearly constant in time when compared to the dynamical behavior of our system. The usual procedure is then to include into the mathematical model a set of parameters which represent this external influence and are supposed to be slowly varying in time. Indeed, a large part of mathematics dealing with the description of (dynamical) systems is at least partly concerned with or motivated by this type of modelling. We mention perturbation theory, bifurcation theory and the theory of structural stability.

The third possibility is to try to really include the connections of the system with the outside world into the description of the system. The system is therefore, so to say, not regarded as an isolated "box," but as a "box" together with the "wires" connecting it to the rest of the world. This third possibility we will call the <u>system theoretic description of a physical system</u>. Of course this goes along with a changing point of view. One does not try to isolate the system "at all costs," but one is especially interested in the continuous interplay of the system and its environment. Since this environment is considered as "unknown," we have to study the set of <u>all</u> dynamical behaviors which can occur at the boundary of the system (the wires of the box), i.e., all behaviors which are compatible with the system under consideration. This whole set is called the <u>external behavior</u> of the system. We should, however, mention that for real systems there may be a very large number of connections with the outside world, whereas in a system-theoretic description we will normally only treat a small number of them and neglect the rest. Hence the same type of questions as arising in the first and second possibility also exists in a system theoretic description. However, we have at least on a conceptual level a way to deal with the influences from and on the outside world. This seems to be an important advantage of the third possibility.

There is another argument in favor of the system theoretic description. In disciplines like physics and chemistry it has been a very successful approach to consider a system as composed of smaller and simpler subsystems which are much easier to describe. Indeed the success story of physics seems to be partly based on its concentration on the study of simple and idealized systems. Afterwards the large real system can then be "understood" in terms of the simple systems which constitute the large system. In fact in celestial mechanics a breakthrough made by Newton, was to consider the solar system as composed of the heavenly bodies, each forming a system on its own, governed by a simple law (Newton's second law), and undergoing forces from the other systems and on its turn exerting forces on them. This approach, called "tearing," gives us the system as a (sometimes complicated) interconnection of all kinds of relatively simple systems. To study the whole system we can study these simple systems separately. But then we should also include in their description their external behavior (i.e., the way in which they can influence and can be influenced by the outside world), since this will be needed in order to determine the behavior of the whole system. The procedure is thus as follows. Tear the system into simple subsystems. Study the systems together with their external behavior. Then interconnect the simple systems again with each other. For example, given an electrical circuit, we can first study the behavior of its elements (capacitors, inductances, resistances, and so on) out of which the circuit is composed. Then by interconnecting these elements in accordance with Kirchhoff's laws one can obtain the original circuit again.

This brings us to another point in favor of the system theoretic approach, which has its roots in technical applications and engineering. Instead of studying the behavior of a complicated system by tearing it, we go the other way around and we want to construct a system with a specified behavior, out of simple building blocks. This leads to the so-called synthesis problem: which building blocks should we use and how should we interconnect them in order to achieve a system with a specified behavior. Clearly to tackle this problem we need a theory of systems which also includes their external behavior.

A more general argument for the system theoretic description, also originating from engineering, has to do with the attitude to consider a system as a device. Usually, this goes together with the so-called input-output framework. One looks at a system as a device which transforms inputs (controls) into outputs. The external behavior of the device is exactly this relationship between input functions and output functions. Clearly, this external behavior of the device is really what counts in applications.

Summarizing, we want to study systems which may be connected with other systems. Therefore, we consider the system as separated from the outside world, but we also incorporate in its description the external behavior of the system. We will assume that this external behavior is given by specifying the possible evolutions in time of a set of variables, which we will call the external variables.

3.1 The notion of state

Apart from connections with other systems there is still another, maybe even more fundamental reason to study the external behavior of a system. This has to do with the notion of *state*. Intuitively the state of a system should contain the whole memory of the system. Knowledge of the system at a certain instance of time, together with the knowledge of all future external influences should totally determine the future (possibly probabilistic) dynamical behavior of the system. Hence, in the case that the system is isolated, the state of the system is all one needs to know in order to predict the future (single) behavior of the system. The usual mathematical structure for this last situation is a set of first-order differential equations in the state variables. Partial differential equations can be seen as first-order differential equations on an infinite-dimensional state space, and many other mathematical descriptions are also variations on this theme.

Of course, this type of modelling presupposes that one knows which variables constitute the state of the system. In many situations, however, a physical system is actually given by a set of "phenomenological" laws, describing the external behavior of the system and not involving the state variables. A simple example is the law for ideal gases $PV = \text{constant}$, which gives the relation between the two external variables P (pressure) and V (volume). A state of the system consists of the positions and velocities of all particles involved. Another simple example is Newton's second law $F = m\ddot{q}$ which is a dynamical compatibility relation between the two external variables F (force) and q (position) as functions of time. The state of this system consists of the position and the velocity, or the position and the momentum. Hence in this case the state can be very easily constructed from the knowledge of the external variables as functions of time, but does not explicitly enter the law $F = m\ddot{q}$. We also consider a (large) electrical network, described by compatibility relations on the voltages and currents on some wires emanating from the network. These compatibility relations do not have to involve the state variables, which are the voltages or currents of (a subset of) the circuit elements _inside_ the network. We see that there can be two reasons for giving the system as a set of compatibility relations ("laws") on the external variables, not involving the state variables:

(i) The state of the system can be very complex, while the external behavior is (relatively) simple.

(ii) The state of the system is not _accessible_ to us; we cannot measure what is going on inside the system.

This second reason goes along with the so-called "black-box" description of a system. We can only observe (or we only care about) what comes into the box and what goes out of it. From an experimental point of view it can be argued that descriptions of physical systems are in first instance always "black box" descriptions.

Concluding, we can say that in many cases the external behavior of a system should be actually taken as the _starting point_ for the description of a system. If we want to know the state of the system we should be able to deduce it from the observations of the external behavior. In system theory this is called the *Realization Problem*: How do we construct from the external behavior

(i) a set of variables which is rich enough to be called the state of the system, and

(ii) the equations governing the evolution of the state?

Since we only want to construct a state which "explains" the external behavior it is of course possible that we end up with a state which does not correspond to the "real physical state" of the system. In the case of a mechanical system we might take instead of the natural state, i.e., the positions and velocities (or momenta) of the particles *another* set of variables which is in one-to-one correspondence with it (notice that we have already mentioned two possibilities for a natural state: positions and velocities, or positions and momenta). For thermodynamic systems it is always possible to find a set of variables which is much smaller than the set of the positions and velocities of all the particles involved, but which on a more axiomatic level can be called the state since it contains all the memory about the external behavior. An extreme example is an ideal gas satisfying PV = constant. This system does not have memory, and hence we do not need a state. The "real physical state" will be *non-minimal*, in the sense that a more parsimonious description of the state may be available. Of course the loss of physical interpretation of the state variables which may occur can be a serious drawback for the theory. In the case of Hamiltonian and gradient systems we will try to combine these notions of a "minimal" and a "physical" state, to end up with a minimal state which is also physically interpretable. The approach which will be taken can be compared with the use of generalized coordinates in classical mechanics.

Finally we remark that we have so far described deterministic systems. In many cases it is of course necessary to take into account uncertainty about the observational data and the parameters of our models, and "identification" will be *See insert p. 17A a central issue. In this context we remark that also in the case of systems with external variables we need a theory which gives information about the validity of our mathematical models, if some parameters are subject to uncertainty (this has much to do with the notion of structural stability).

We now describe a Hamiltonian system from this point of view.

Consider a point mass m with position q_1, influenced by a force F_1. According to Newton's second law, the relation between q_1 and F_1 as functions of time is given by

$$(3.1) \qquad m\ddot{q}_1 = F_1$$

Note that we see F_1 as a basic variable and that (3.1) expresses a compatibility relation between forces and positions. Hence we have an external (linear) system

$$(3.2) \qquad \Sigma_e := \{(q_1(\cdot),F_1(\cdot)):\mathbb{R} \to \mathbb{R}^2 \ (q_1(\cdot),F_1(\cdot)) \in \mathscr{L}^1_{loc} \text{ and } m\ddot{q}_1 = F_1, \text{ with equality in the sense of distributions}\}$$

The notion of state, if we assume a probabilistic description of the uncertainties, is that of a conditional probability density, given the observations, and is in general infinite-dimensional.

In fact Σ_e is an external input-output system with input $u_1 = F_1$. A minimal realization of Σ_e is given by

$$(3.3) \qquad \begin{aligned} \dot{q}_1 &= \frac{1}{m} p_1 \\ \dot{p}_1 &= u_1 \end{aligned} \qquad , \; y_1 = q_1$$

i.e., a linear input-output system $\Sigma(A,B,C)$: $\frac{dx}{dt} = Ax(t)+Bu(t)$ with $A = \begin{pmatrix} 0 & \frac{1}{m} \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $C = (1\ 0)$. Any definition of a Hamiltonian system surely ought to include systems (3.2) and (3.3). The basic observation is that the state space (q_1,p_1) can be seen as a symplectic space with the usual symplectic form $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Then A as above is a Hamiltonian matrix, i.e., A satisfies $A^TJ + JA = 0$, and B and C are related as $B^TJ = c$. Furthermore we notice that the space of inputs and outputs (y_1,u_1) can be also seen as a symplectic space with the symplectic form $J^e = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

Next we look at another mechanical system. Consider a particle attached to a spring with spring constant k. Assume that we can control the position q_2 of the particle. We take as output the force F_2 exerted by the spring on the particle, i.e., the force that we experience if we control the particle in a certain position. This yields the static system $F_2 = -kq_2$, which can be also written as

$$(3.4) \qquad F_2 = - \frac{dV}{dq_2}(q_2)$$

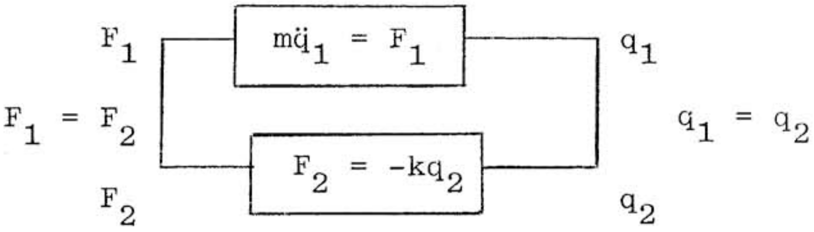
with $V(q_2) = \frac{1}{2}kq_2{}^2$ the potential energy. We regard (3.4) as a static Hamiltonian system with input q_2 and output F_2.

Equation (3.4) defines a Lagrangian submanifold in the (q_2,F_2)-space with generating function $V(q_2)$. Instead of the potential energy $\frac{1}{2}kq_2{}^2$ corresponding to a linear spring we can take an arbitrary potential energy function $V(q_2)$. Notice also that (3.4) is an example where external forces are not necessarily inputs.

Finally we can <u>interconnect</u> the Hamiltonian systems (3.3) and (3.4) by setting

$$(3.5) \qquad q_1 = q_2 \quad , \quad F_1 = F_2$$

(this can be regarded as Newton's third law) :



The interconnection (3.5) is a particularly simple example of what one might call a Hamiltonian interconnection. The system resulting from the interconnection has the form (setting $q=q_1=q_2$):

$$(3.6) \qquad mq + kq = 0, \text{ or }, \frac{d}{dt}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{m} \\ -k & 0 \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix}$$

This constitutes a Hamiltonian vectorfield, or as we shall say an autonomous (i.e., no inputs) Hamiltonian system. As outputs we could take the position q, or the position q together with $-\frac{dV}{dq}(q) = -kq$, which is now the <u>internal</u> force.

4. Optimization, Complexity and Monte-Carlo Techniques(10)

Since for systems with external variables we are dealing with synthesis of possibly new systems, it is necessary to evaluate in what sense the synthesis is successful. This can be done using ad-hoc qualitative criteria or may be formalized by formulating the notion of a best synthesis. Mathematically, this means defining the set of feasible solutions to the synthesis problem by specifying the constraints (for example, static and dynamical laws, constraints on the resources, constraints on the inputs and outputs) and by specifying a performance function (usually real-valued) which is optimized (minimized or maximized). Even though the best synthesis cannot be implemented (for example, due to economic considerations), this formulation allows us to compare different syntheses. The actual optimization procedure will be carried out by devising algorithms which will then be implemented on a computing machine. A natural question then is: how complex is this computation? In section 4.1, we discuss some of these ideas in the context of combinatorial optimization. This is a striking example of the interaction between ideas of systems and computer science. Indeed, understanding and dealing with complexity is one of the outstanding scientific problems of modern technological systems.

4.1 Combinatorial Optimization and Complexity

An instance of an optimization problem is a pair $(F,\mathscr{X})$ where $\mathscr{X}$ is a set and F the cost function is a function

$$F \; : \; \mathscr{X} \to \mathbb{R}^1$$

where $\mathbb{R}^1$ is the set of real numbers. The problem is to find $\hat{x} \in \mathscr{X}$ such that

$$F(x) \leq F(x) \quad \forall\, x \in \mathscr{X}.$$

An optimization is a set I of instances of an optimization problem. In an instance, we are given the "input data" and have enough information to obtain a solution. A problem is a collection of instances, usually all generated in a similar way.

Example 1: Travelling Salesman Problem (TSP)

In an instance of the TSP we are given an integer $n > 0$ and the distance between any pair of n cities in the form of a $n \times n$ matrix (d_{ij}), $d_{ij} \in Z^+$ (the set of positive integers). A tour is a closed path that visits every city exactly once. The problem is to find a tour of minimal length. If we denote by

$$\mathscr{X} = \{\text{all cyclic permutations } \pi \text{ on n objects}\} \quad ,$$

then a cyclic permutation represents a tour if we interpret $\pi(j)$ to be the city visited after city j, j=1, 2, ... n. Then the cost function

$$F(\pi) = \sum_{j=1}^{n} d_{j\pi(j)} \quad .$$

Example 2: (Minimal Spanning Tree (MST))

A spanning tree is an undirected graph (V,E) (where V =

set of nodes and E = set of edges), that is connected and acyclic.

We are given an integer $n > 0$ and an $n \times n$ symmetric matrix (d_{ij}), $d_{ij} \in Z^+$. Let

$$\mathscr{X} = \{\text{all spanning trees } (V,E),\ V = \{1,2,\ \ldots n\}\}$$

Then the MST problem is to minimize

$$\begin{aligned} F \ &: \ \mathscr{X} \to \mathbb{R}^1 \\ &; \ (V,E) \mapsto \sum_{i,j)\in E} d_{ij} \quad . \end{aligned}$$

The above are examples of combinatorial optimization problems.

In the modern theory of computation, we are interested in algorithms which are efficient in the sense that the number of steps required to solve the problem grows as a polynomial in the size of the input. MST is an example of a problem which is efficient in the above sense, while TSP is an example of a problem for which no efficient algorithm is known, and generally one has to resort to heuristics. These notions can be made more precise by defining two classes of problems, P and NP. We do this now.

We assume that the set $\mathscr{X}$ and the function F are given in terms of two algorithms, $A_{\mathscr{X}}$ and A_F. The algorithms $A_{\mathscr{X}}$, given x and a set S of parameters will decide whether $x \in \mathscr{X}$, the feasible set. On the other hand A_F, given a feasible x and another set of parameters Q, returns the value F(x). An instance of an optimization problem is a representation of the parameters in S and Q, using a fixed, finite alphabet and an appropriate coding. We now define a recognition version of the optimization problem:

Given a representation of S and Q and an integer L, is there a feasible solution such that $F(x) \leq L$?

This problem has a yes and no answer. We denote by the class P, the class of recognition problems that can be solved by a polynomial-time algorithm. This can be given a precise definition in terms of, say, a Turing machine. It turns out that this class has a remarkable property, namely, if it can be solved in polynomial time by one model of computation then it can be solved in polynomial time by all reasonable models of computation.

There is another class of recognition problems, the class NP. For a problem to be in NP, we do not require that every instance can be answered in polynomial time by some algorithm. We simply require that if x is a yes instance of the problem, then there exists a concise certificate (of length bounded by a polynomial in the size of x) of x, which can be checked in polynomial time for validity.

It is easy to see that $P \subset NP$. It is unknown if $P = NP$. The travelling salesman problem is in NP and it is notoriously difficult. If P were equal to NP then the travelling salesman problem would have a polynomial time algorithm.

We say that a recognition problem A_1 polynomially transforms to another recognition problem A_2 is given any string x we can construct a string y, within polynomial in length of x time, such that x is a yes instance of A_1 iff y is a yes instance of A_2.

A recognition problem $A \in NP$ is said to be NP-complete if all other problems in NP polynomially transform to A. Combining the conjecture $P \subset NP$ and the definition of NP-completeness we see that characterizing a combinatorial problem to be difficult means showing it is NP-complete.

4.2 A Probabilistic Algorithm for Combinatorial Optimization(11)

Simulated annealing, as proposed by Kirkpatrick, is a recent Monte-Carlo algorithm for combinatorial optimization. Simulated annealing is a variation on an algorithm introduced by Metropolis for approximate computation of mean values of various statistical-mechanical quantities for a physical system in equilibrium at a given temperature. In simulated annealing the temperature of the system is slowly decreased to zero; if the temperature is decreased slowly enough the system should end up among the minimum energy states or at least among states of sufficiently low energy. Hence the annealing algorithm can be viewed as minimizing a cost function (energy) over a finite set (the system's states). Simulated annealing has been applied to several combinatorial optimization problems including the traveling salesman problem, computer design problems, and image reconstruction problems with apparently good results.

The annealing algorithm consists of simulating a nonstationary finite-state Markov chain which we shall call the annealing chain. We now describe the precise relationship between this chain and the finite optimization problem to be solved. Here and in the sequel we shall take $\mathbb{R}$ to be the real numbers, $\mathbb{N}$ the natural numbers, and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, and we shall denote by $|A|$ the cardinality of a finite set A. Let Ω be a finite set, say $\Omega = \{1, \ldots, |\Omega|\}$, and $U_i \in \mathbb{R}$ for $i \in \Omega$; we want to minimize U_i over $i \in \Omega$. Let $T_k > 0$ for $k \in \mathbb{N}_0$. Ω shall be the state-space for the annealing chain and we shall refer to $\{U_i\}_{i\in\Omega}$ as the <u>energy function</u> and $\{T_k\}_{k\in\Omega}$ as the <u>annealing schedule of temperatures</u>. Let $\pi^{(k)} = [\pi_i^{(k)}]_{i\in\Omega}$ (a row vector) be a Gibbs distribution over the energies $\{U_i\}_{i\in\Omega}$ at a temperature T_k, i.e.,

$$\pi_i^{(k)} = \frac{e^{-U_i/T_k}}{\sum_{j\in\Omega} e^{-U_j/T_k}} , \qquad i \in \Omega ,$$

for all $k \in \mathbb{N}_0$. The annealing chain will be constructed such that at each time k the chain has $\pi^{(k)}$ as its unique invariant distribution, i.e., at each time k the annealing chain shall have a 1-step transition matrix $p^{(k,k+1)} = [p_{ij}^{(k,k+1)}]_{i,j\in\Omega}$ such that $\pi = \pi^{(k)}$ is the unique solution of the vector equation $\pi = \pi P^{(k,k+1)}$. The motivation for this is as follows. Let S^* be the minimum energy states in Ω. Now if $T_k \to 0$ as $k \to \infty$ then

$$\pi_i^{(k)} \to \begin{cases} \frac{1}{|S^*|} & \text{if } i \in S^* , \\ 0 & \text{if } i \notin S^* , \end{cases}$$

as $k \to \infty$, i.e., the invariant distributions converge to a uniform distribution over the minimum energy states. The hope is then that the chain itself converges to the minimum energy states.

We now show how Metropolis constructs a transition matrix $p^{(k,k+1)}$ with invariant vector $\pi^{(k)}$ for $k \in \mathbb{N}_0$. Let $Q = [q_{ij}]_{i,j\in\Omega}$ be a symmetric and irreducible stochastic matrix, and let

$$p_{ij}^{(k,k+1)} = \begin{cases} q_{i,j}e^{-(U_j-U_i)/T_k} & \text{if } U_j \quad U_i \ , \\ q_{ij} & \text{if } U_j \le U_i,\ j \ne i, \\ 1 - \sum\limits_{\ell\ne i} p_{i\ell}^{(k,k+1)} & \text{if } j = 1 \ , \end{cases}$$

for all $i,j \in \Omega$ and $k \in \mathbb{N}_0$. Then it is easily verified that $\pi^{(k)} = \pi^{(k)}p^{(k,k+1)}$ for all $k \in \mathbb{N}_0$. In fact, $p^{(k,k+1)}$ and $\pi^{(k)}$ satisfy the reversibility condition

$$p_{ji}^{(k,k+1)}\ \pi_j^{(k)} = \pi_i^{(k)}p_{ij}^{(k,k+1)} \ , \qquad i,j \in \Omega,$$

for all $k \in \mathbb{N}_0$. Let $\{x_k\}_{k\in\mathbb{N}_0}$ be the annealing chain with 1-step transition matrices $\{P^{(k,k+1)}\}_{k\in\mathbb{N}_0}$ and some initial distribution, constructed on a suitable probability space (M,Λ,P). Let $p_i^{(k)} = P\{x_k = i\}$ for $i\in\Omega$ and $k\in\mathbb{N}_0$.

The annealing chain is simulated as follows. Suppose $x_k = i \in \Omega$. Then generate a random variable $y \in \Omega$ with $P\{y = j\} = q_{ij}$ for $j \in \Omega$. Suppose $y = j \in \Omega$. Then set

$$x_{k+1} = \begin{cases} j & \text{if } U_j \le U_i, \\ j & \text{if } U_j > U_i, \text{ with probability } e^{-(U_j-U_i)/T_k} \ , \\ i & \text{else.} \end{cases}$$

Hence, we may think of the annealing algorithm as a "probabilistic descent" algorithm where the Q matrix represents some prior distribution of "directions," transitions to same or lower energy states are always allowed, and transitions to higher energy states are allowed with positive probability which tends to 0 as $k \to \infty$ (when $T_k \to 0$ as $k \to \infty$). This algorithm is a striking example of the power of thinking by analogy, in this case analogy with problems in statistical mechanics, notably the theory of spin glasses.

4.3 Dynamic Programming(12)

The optimization problems we have described so far are static in the sense that we are concerned only with equilibrium situations where the systems under consideration do not change with time. The more general and realistic situation that we need to consider is where there is a dynamical (possibly probabilistic) description of the system and where we have imperfect partial observations of the state of the system and on the basis of these observations we are required to use feedback control to optimally control the system according to some pre-assigned performance criterion. The general method for doing this is dynamic programming and we illustrate this by considering a deterministic dynamical situation.

Let us suppose that the state of the system evolves according to a differential equation

$$\frac{dx}{dt} = f(x(t), u(t), t) \; ; \; x(0) = x$$

where $x(t) \in \mathbb{R}$ is the state of the system, $t \to u(t) : [0,\infty) \to \mathbb{R}$ is the control or decision function and we are required to choose this function, in some appropriate class to minimize the performance function

$$J(u;x) = \int_0^T L(x(t),u(t))dt \ .$$

A minimizing control is called optimal. We recognize this as a problem in the Calculus of Variations. The philosophy of dynamic programming is to embed the problem in a family of problems:

$$(4.1) \qquad \begin{cases} \frac{dx}{dt} = f(x(t),\ u(t),t) \\ x(x) = x \quad , \end{cases}$$

where the initial time in arbitrary and the initial state $x \in \mathbb{R}$ is arbitrary and we choose the control function: $t \to u(t) : [s,\infty) \to \mathbb{R}$ to minimize

$$(4.2) \qquad J(u;s,x) = \int_s^T L(x(t),u(t)t)dt \quad .$$

The idea of dynamic programming is based on the principle of optimality which states: An optimal policy has the property that whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

Let us apply this to the variational problem at hand. We denote by

$$V(s,x) = \operatorname*{Inf}_{u} \; J(u;s,x)$$

Let us assume that this minimum is obtained for a control function $t \rightarrow u^*(t)$ and let $x^*(t)$ be the corresponding solution of the differential equation (4.1). Let us apply the control u over the interval $(s,s + \Delta s)$ and the optimal control $u^*(\cdot)$ for the reamining time. Then by the Principle of Optimality

$$V(s,x) = \operatorname*{Min}_{u}[\int_{s}^{s+\Delta} L(x(t),u(t),t)dt + V(s+\Delta s,x(s)+\Delta x(s))]$$

Now if Δ is small, doing a first-order Taylor approximation, we obtain:

$$V(s,x) = \operatorname*{Min}_{u}[L(x(s),u(s)s)\Delta + V(s,x(s)) + \frac{\partial V}{\partial x}\Delta x + \frac{\partial V}{\partial s}]$$

and hence when $\Delta \rightarrow 0$, we get the partial differential equation

$$-\frac{\partial V}{\partial s} = \operatorname*{Min}_{u}[L(x(s),u(s)s) + \frac{\partial V}{\partial x} f\ (x(s),u(s),s)] \quad .$$

This partial differential equation is nothing but the Hamiltonian Jacobi equation of Classical Mechanics. We however see that by carrying out the minimization within square brackets we obtain the optimal control function as a function of the state of the system, which is the idea behind feedback control.

The philosophy and method of dynamic programming is extremely general and is probably the only general method for dealing with dynamic decision problems, even in the presence of uncertainties.

5. From Data to Models[13]

We have discussed in the introduction that one of the new problems in the theory of dynamical systems with external variables is the building of mathematical models from observed data, when there may not be any underlying physical law to obtain on a priori mathematical structure of the model.

We take a possible logical view of this model-building exercise to mean the removing of redundancies in the data and thus to discover regular statistical features. We take as a measure of determination of the best mathematical the shortest length with which this data can be described, say in binary digits. Following Rissanen we call this stochastic complexity of the model, in a preassigned class of models, and the criterion for determining the model is called the Minimum Description Length criterion. These ideas are inspired by the algorithms notion of information due to Kolmogoroff and Chaitin.

Mathematical models are also used for prediction purposes and hence a measure which codifies both the representation of the data and the accuracy with which prediction can be made is desirable. This leads to Rissanen's Predictive Maximum Description Length Principle which we now describe.

The probabilistic models we consider consist of indexed densities $f_\alpha(x|u)$, or ultimately probabilities $P_\alpha(x|u)$, where x_1, ..., x_n, also written as x^n, denotes a sample of length n as a response or "output" to another "input" sample

$u = u^n$ of the same length. Because the input sample adds nothing new in principle, we drop it to simplify the notations. We let the lower case letters denote both random variables and their values, letting the context tell which is meant. The data items are often numerical, but, of course, not always. When numerical, each number in the binary notation, say, has only some number r of fractional digits. Hence, when the model is a density it assigns a probability to x, which is obtained by integrating the density over the n-dimensional cube of edge length 2^{-r} with x as the center. We denote this induced probability function by $P_\alpha(x)$ without indicating the implicitly understood precision r, which we otherwise do not need. The index α may be taken sufficiently general to allow comparision of nested and non-nested models alike. However, it is the number of parameters that turns out to be interesting quantity, and we take for simplicity the index to be of the form $\alpha = (k,\theta)$, where k denotes the number of components in the parameter vector $\theta = (\theta_1,\ldots,\theta_k)$, and $k = 0,1,\ldots$. The value $k = 0$ corresponds to the empty parameter λ.

We are interested in predicting the sequence x as well as coding it. The former may be viewed as a special predictive form of coding, and we gain generality by proceeding with the coding interpretation. Often, we wish to model the data such that the individual observations are independent. Then, instead of coding a sequence the relevant problem is to consider coding of the n - element unordered set $\{x_i\}$, where repeated occurrences of a value are preserved. The required modification for such a case will be discussed below. Predictive coding means that we model the conditional density for the possible values of the "next" observation x_{i+1} thus

$$(5.1) \qquad f_{k,\hat{\theta}(t)}(x_{t+1} \mid x^t) \ ,$$

where $\hat{\theta}(t) = \hat{\theta}(x^t)$ is an estimation algorithm for the parameter θ with k components. Such a density allows us to encode the observation x_{i+1} to the precision r with the "ideal" code length $- \log P_{k,\hat{\theta}(t)}(x_{t+1}|x^t)$, which, as just explained, is represented by $- \log f_{k,\hat{\theta}(t)}(x_{t+1}|x^t)$. The word "ideal" means that if the possible values of the next observation indeed are distributed as modeled, then no prefix code exists with a shorter mean length. Whenever we wish to express the code length as the number of binary digits in the coding string, the logarithm is to be taken to the base 2; otherwise, its base does not matter. By adding all these ideal code lengths, we get the total code length

$$(5.2) \qquad L(x|k) = - \sum_{t=0}^{n-1} \log f_{k,\hat{\theta}(t)}(x_{t+1}|x^t) \ .$$

This may be minimized with respect to k to give the estimator $\hat{k}(n) = \hat{k}(x^n)$, which with the last data point defines the final estimate $\hat{\theta}(n)$ having $\hat{k}(n)$ components.

How should we select the estimate $\hat{\theta}(t)$ for each k? On first thought one might think of picking it so as to minimize the ideal code length $- \log f_{k,\theta}(x_{t+1}|x^t)$, which amounts to the maximum likelihood estimator. But, clearly, this cannot be done, because such a minimization would make $\hat{\theta}(t)$ a function of x_{t+1}, which, in turn, would make decoding impossible. Indeed, decoding of x_{t+1} requires the knowledge of $\hat{\theta}(t)$, which therefore must not depend on the value x_{t+1} to be decoded. We are faced with the central issue in inductive inference, and we reason as follows: In the light of past observations the best single value of the parameter for encoding the "next" observations, x_{i+1}, $i=0,1,\ldots,t-1$ is the value that minimizes the sum $- \sum_{i=0}^{t-1} \log f_{k,\theta}(x_{i+1}|x^i)$. This is the maximum likelihood estimate $\hat{\theta}(t)$, except that we add the restriction that the predicted density (5.1) is positive for every possible value of x_{t+1}, which is required to make (5.2) meaningful for all data sequences. We might then say that this choice for the estimator $\hat{\theta}(t)$ is based upon the hope that the predicted distribution (5.1) for the new observation x_{t+1} is like it was in the past.

The minimization of (5.2) requires the initial estimate $\hat{\theta}(0)$ for each number of components k. The traditional way to calculate such is to select more or less arbitrarily a priori density function for the parameters and then take one of the maximizing values as the estimate $\hat{\theta}(0)$. The predictive approach, however, offers a different way, and one which avoids the both conceptually and technically difficult problem of specifying the prior densities. Indeed, what (5.2) really requires is the specification of a density function $f(x_1)$ for the first observation such that it reflects our prior knowledge about its value. Technically, we may take this density function to be in the parametric family and specified by the empty parameter λ. Such a distribution is often much easier to pick than a priori for the parameters. For example, if the prior knowledge consists of the fact that the set of possible values of x_1 is finite, M, put $- \log f(x_1) = \log M$. The procedure to compute (5.2) for each selected number of parameters k is then as follows: The first observation x_1 is encoded with the ideal code length $- \log f(x_1)$, where the density is selected to represent our knowledge, often ignorance, about the value x_1. We continue encoding the next observation with this same density until one parameter can be uniquely fitted, and we increase the number of fitted parameters in this manner one by one until the set value k, needed in the evaluation of (5.2), is reached.

The minimized code length (5.2) does not quite represent the complexity of the sequence x, because it is conditioned on the optimizing number of parameters, which clearly is required in the decoding process. This value can be given in a coded form as a preamble in the entire code string. Because the decoder will have to be able to separate the binary codeword representing $\hat{k}(n)$ from the subsequent code of the data without a separating comma, the preamble must be a so-called prefix code. From information-theoretic considerations one can show that encoding the natural number k by a prefix code requires

$$(5.3) \qquad L^*(k) = \log^* k + \log c$$

binary digits, where $\log^* k = \log k + \log\log k + \ldots$, the sume including all the positive iterates, and c is the constant, about 2.865, that makes $\sum_{n=1}^{\infty} 2^{-L(n)} = 1$. Therefore, we may define the (semi) <u>predictive complexity</u> of the sequence x, relative to the selected class of models as

$$(5.4) \qquad \ell_{SP}(x) = \min_k \{ L(x|k) + \log^* k + c \} .$$

The word "semi" suggests that the optimizing number of parameters, which we still write as k(n), is not determined the predictive way. Nevertheless, what is important is that this complexity defines a proper density, which is proportional to $h_{SP}(x) = 2^{-\ell(x)}$. In fact,

$$\int_{x\in X^n} h_{SP}(x)dx = \sum_{k=0}^{\infty} 2^{-L^*(k)} \int_{x\in X^n} 2^{-L(x|k)} dx \leq 1,$$

where the integrals are taken over all strings of length n. To avoid misunderstandings we emphasize that the main effect for penalizing the number of parameters in (5.4) is by no means due to the second term, $\log^* k$. In fact, in most if not all the cases the minimizations of 5.2, where no such term appears, and (5.4) produce exactly the same number of components, which is why we may safely use the same symbol to denote both.

We can apply the above discussed inductive reasoning to obtain a purely predictive complexity. Indeed, let $\hat{k}(t)$ denote the minimizing number of parameters in (5.2), where n is replaced by t. Then we may regard the pair $(\hat{k}(t), \hat{\theta}(t))$ to represent our best estimate of the conditional density for the possible values of the "next" observation x_{t+1} available at time t. Adding the resulting ideal code lengths we get the purely <u>predictive complexity</u> as follows

$$\ell_p(x) = - \sum_{t=0}^{n-1} f_{\hat{k}(t), \hat{\theta}(t)}(x_{t+1} | x^t), \tag{5.5}$$

where $\hat{k}(0) = 0$ and $\hat{\theta}(0) = \lambda$, representing the empty set of parameters. In other words, the initial density $f(x_1)$ is determined as described above. The predictive complexity defines a density $h_p(x) = 2^{-\ell p(x)}$, which is proper in that it integrates to unity over the sequences of the same length.

We next study to what extent prediction error measures can be interpreted as code lengths, which at the same time illustrates how the large classes of models as studied here are typically generated. Let $\hat{x}_{t+1} = g_\theta(x^t)$ denote a parametric predictor of x_{t+1}, where the parameter is to be determined from the past data. Usually, the predictors are defined by recurrence equations such as of the ARMA type. One may view this process as a means of accounting for the dependencies in the data, which when done well causes the prediction errors to be nearly independent. Next, let $\delta_t(x_{t+1}, \hat{x}_{t+1})$ denote a measure of the prediction error.

Now define

$$(5.6) \qquad f_{t,\theta}(x_{t+1}|x^t) = K(x^t,\theta)2^{-\delta t(x_{t+1},\hat{x}_{t+1})} ,$$

where $K(x^t,\theta)$ denotes that number for which $f_{t,\theta}(y|x^t)dy = 1$. We then see that

$$- \log f_{t,\theta} (x_{t+1}|x^t) = \delta_t(x_{t+1}, \hat{x}_{t+1}) - \log K(x^t,\theta)$$

represents an ideal code length for the observation x_{t+1} given the past data. With a suitable estimator $\hat{\theta}(t) = \hat{\theta}(x^t)$ the total ideal code length takes the form

$$(5.7) \qquad L(x|k) = \sum_{t=0}^{n-1} \delta_t(x_{t+1},\hat{x}_{t+1}) - \sum_{t=0}^{n-1} \log K(x^t, \hat{\theta}(t)).$$

We see that this predictive MDL criterion differs from the first sum, involving the prediction errors, only to the extent the second term depends on k. Most of the usual prediction error measures actually depend only on the difference $e_{t+1} = x_{t+1} - \hat{x}_{t+1}$, and, moreover, often the possible values of x_{t+1} range from $-\infty$ to $+\infty$. Then we see that $K(x^t, \hat{\theta}(t)) = K(x^t)$, and the difference between the two criteria amounts to a constant.

Notes and References

(1) See for example, Adolf Grünbaum, *The Foundations of Psychoanalysis*, U. of Calif. Press, Berkeley, 1985.

(2) Eugene P. Wigner:

(3) Edmund Husserl, *The Crisis of European Sciences*, Northwestern University Press, Evanston, 1970.

(4) For a striking example of the interaction between pure mathematics and physics, see, M.F. Aliyah, *Geometry of Yang-Millo Fields*, Lectures at Scuola Normale Superiore, Pisa, 1978.

(5) For an elegant study of this viewpoint, see V. Arnold: *Mathematical Methods of Classical Mechanics*, Springer, New York, 1978, and R.A. Abraham and J.E. Marsden; *Foundations of Mechanics*, Benjamin/Cummings, Reading, MA, 1978.

(6) See D. Ruelle: *Thermodynamic Formalism*, Addison-Wesley, Reading, Mass., 1978.

(7) For a more detailed discussion, see H. Dym and H.P. McKean: *Fourier Series and Integrals*, Academic Press, New York, 1972.

(8) The study of linear dynamical systems is today a big subject. The interested reader should consult; R.E. Kalman, P.L. Falb, and M.A. Arbib: *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969. For Grothendieck's theorem from the viewpoint of linear system theory, see M. Hazewinkel and C.F. Martin: A

Short Elementary Proof of Grothendieck's Theorem on algebraic vector bundles over the projective line, J. of Pure and Applied Algebra, 25, pp. 207-211, 1982. The invariants discussed here are Kronecker invariants. See F.R. Gantmacher: The Theory of Matrices, Chelsea, New York, 1959.

(9) The ideas presented in this section are due to J.C. Willems and A.J. Van der Schaft. See, J.C. Willems: The Analysis of Physical Systems, Ricerche di Automatica, Special Issue on System Theory and Physics, Vol. X, No. 2, pp. 71-106, 1979.

(10) For an account of combinatorial optimization and related questions of computational complexity, see C.H. Papadimitriou:

on which this section is based.

(11) The idea of simulated annealing was proposed by Kirkpatrick. See S. Kirkpatrick, D.C. Gelalt and M.P. Vecchi: Optimization by Simulated Annealing, Science, 220, pp. 621-680, 1983. The reference to Metropolis is: N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller: Equations of State Calculations by Fast Computing Machines, J. of Chem. Phys., 21, pp. 1087-1091, 1953.

(12) For a lucid description of these ideas, see R. Bellman: Adaptive Control Processes, Academic Press, New York 1960.

(13) The ideas expressed in this section are due to J. Rissanen. See for example, Stochastic Complexity and

Predictive Modelling, to appear in Annals of Statistics. The works of Kolmogoroff and Chaitin related to the algorithms notion of information are: A.N. Kolmogoroff, Three Approaches to the Quantitative Definition of Information, Problems of Information Transmission, 1, pp. 4-7, 1965. J.G. Chaitin, A Theory of Program Size Formally Identical to Information Theory, J. ACM, 22, pp. 329-340, 1975.