Chapter 1
Introduction

1 SYSTEMS AND CYBERNETICS

Civilization is rapidly becoming very dependent on large-scale systems of men, machines, and environment. Because such systems are often unpredictable, we must rapidly develop a more sophisticated understanding of them to prevent serious consequences. Very often the ability of the system to carry out its function (or alternatively, its catastrophically failing to function) is a property of the system as a whole and not of any particular component. The single most important rule in the management of large scale systems is that one must account for the entire system - the sum of all the parts. This most likely involves the discipline of "differential games." It is reasonable to predict that cybernetic methods will be relevant to the solution of the greatest problems that face man today.

Cybernetics is the science of communication and control in machines and living creatures [133]. Nature employs the best cybernetic systems that can be conceived. In the neurological domain of living beings, the ecological balance involving environmental feedback, the control of planetary movements, or the regulation of the temparature of the human body, the cybernetic systems of nature are fascinating in their accuracy and efficiency. They are cohesive, self-regulating and stable systems; yet they do have the remarkable adaptability to change and the inherent capacity to use experience of feedback to aid the learning process.

Sustained performance of any system requires regulation and control. In complicated machinery the principles of servomechanism and feedback control have long been in effective use. The control principles in cybernetics are the error-actuated feedback and homeostasis. Take the case of a person driving a car. He keeps to the desired position on the road by constantly checking the deviation through visual comparison. He then corrects the error by making compensating movements of the steering wheel. Error sensing and feedback are both achieved by the driver's brain which coordinates his sight and muscular action. Homeostasis is the self-adjusting property that all living organisms possess and that makes use of feedback from the environment to adjust metabolism to changing environmental conditions. Keeping the temperature of the human body constant is a good example of homeostasia.

The application of cybernetics to an environmental situation is much more involved than the servomechanism actuating "feedback correction." The number of variables activating in the system are plentiful. The variables behave in stochastic manner and interactive relationships among them are very complex. Examples of such systems in nature are meteorological and environmental systems, agricultural crops, river flows, demographic systems, pollution, and so on. According to complexity of interactions with various influences in nature, these are called cybernetical systems. Changes take place in a slow and steady manner, and any suddenness of change cannot be easily perceived. If these systems are not studied continuously by using sophisticated techniques and if predictions of changes are not allowed to accumulate, sooner or later the situation is bound to get out of hand.

The tasks of engineering cybernetics (self-organization modeling, identification, optimal control, pattern recognition, etc.) require development of special theories which, although look different, have many things in common. The commonality among theories that form the basis of complex problem-solving has increased, indicating the maturity of cybernetics as a branch of science [37]. This leads to a common theory of self-organization modeling that is a combination of the deductive and inductive methods and allows one to solve complex problems. The mathematical foundations of such a common theory might be the approach that utilizes the black box concept as a study of input and output, the neural approach that utilizes the concept of threshold logic and connectionism, the inductive approach that utilizes the concept of inductive mechanism for maintaining the composite control of the system, the probabilistic approach that utilizes multiplicative functions of the hierarchical theory of statistical decisions, and Godel's mathematical logic approach (incompleteness theorem) that utilizes the principle of "external complement" as a selection criterion.

The following are definitions of terms that are commonly used in cybernetic literature and the concept of black box.

1.1 Definitions

1. A *system* is a collection of interacting, diverse elements that function (communicate) within a specified environment to process information to achieve one or more desired objectives. Feedback is essential, some of its inputs may be stochastic and a part of its environment may be competitive.
2. The *environment* is the set of variables that affects the system but is not controlled by it.
3. A *complex system* has five or more internal and nonlinear feedback loops.
4. In a *dynamic system* the variables or their interactions are functions of time.
5. An *adaptive system* continues to achieve its objectives in the face of a changing environment or deterioration in the performance of its elements.
6. The rules of behavior of a *self-organizing system* are determined internally but modified by environmental inputs.
7. *Dynamic stability* means that all time derivatives are controlled.
8. A *cybernetic system* is complex, dynamic, and adaptive. Compromise (optimal) control achieves dynamic stability.
9. A *real culture* is a complex, dynamic, adaptive, self-organizing system with human elements and compromise control. Man is in the feedback loop.
10. A *cybernetic culture* is a cybernetic system with internal rules, human elements, man in the feedback loop, and varying, competing values.
11. *Utopia* is a system with human elements and man in the feedback loop.

The characteristics of various systems are summarized in Table 1.1, where 1 represents "always present" and a blank space represents "generally absent." The differences among the characteristics of Utopia and cybernetic culture are given in Table 1.2.

Table 1.1. Characteristics of various systems

Characteristics	System	Complex system	Dynamic system	Adaptive system	Self-organizing system	Cybernetic system	Real culture	Cybernetic culture	Utopia
Collection of interacting, diverse elements, process information, specified environment, goals feedback	1	1	1	1	1	1	1	1	1
At least five internal and nonlinear feedback loops		1				1	1	1	
Variables and interactive functions of time			1			1	1	1	
Changing environment deteriorating elements				1		1	1	1	
Internal rules					1		1	1	
Compromise (Optimal) control						1 1	1 1	1 1	
Human elements							1	1	1
Dynamic stability						1		1	
Man in feedback loop						1	1	1	1
Values varying in time and competing						1		1	

Table 1.2. Differences among Utopia and cybernetic culture

Characteristic	Utopia	Cybernetic culture
Size	Small	Large
Complex	No	Yes
Environment	Static, imaginery	Changing, real
Elements deteriorate	No	Yes
Rules of behavior	External	Internal
Control	Suboptimized	Compromised
Stability	Static	Dynamic
Values	Fixed	Varying, Competing
Experimentation	None	Evolutionary operation

1.2 Model and simulation

Let us clarify the meaning of the words *model* and *simulation.* At some stage a *model* may have been some sort of small physical system that paralleled the action of a large system; at some later stage, it may have been a verbal description of that system, and at a still later - and hopefully more advanced - stage, it may have consisted of mathematical equations that somehow described the behavior of the system.

A model enables us to study the various functions and the behavioral characteristics of a system and its subsystems as well as how the system responds to given changes in inputs or reacts to changes in parameters or component characteristics. It enables us to study the extent to which outputs are directly related to changes in inputs - whether the system tends to return to the initial conditions of a steady state after it has been disturbed in some way, or whether it continues to oscillate between the control limits. A cybernetic model can help us to understand which behavior is relevant to or to what extent the system is responsible for changes in environmental factors.

Simulation is a numerical technique for conducting experiments with mathematical and logical models that describe the behavior of a system on a computer over extended periods of time with the aim of long-term prediction, planning, or decision-making in systems studies. The most convenient form of description is based on the use of the finite-difference form of equations.

Experts in the field of simulation bear a great responsibility, since many complex problems of modern society can be properly solved only with the aid of simulation. Some of these problems are economic in nature. Let us mention here models of inflation and of the growing disparity between rich and poor countries, demographic models, models for increased food production, and many others. Among the ecological problems, primary place is occupied by problems of environmental pollution, agricultural crops, water reservoirs, fishing, etc. It is well known that mathematical models, with the connected quantities that are amenable to measurement and formalization, play very important roles in describing any process or system. The questions solved and the difficulties encountered during the simulation complex systems modeling are clearly dealt with in this book.

It is possible to distinguish three principal stages of the development of simulation:

$$\begin{pmatrix}\text{Experts without}\\ \text{computers}\end{pmatrix} \to \begin{pmatrix}\text{Man-machine}\\ \text{Dialogue systems}\end{pmatrix} \to \begin{pmatrix}\text{Computer without}\\ \text{experts}\end{pmatrix}$$

We are still at the first stage; man-machine dialogue systems are hardly used at this time. Predictions are realized in the form of two or three volumes of data tables compiled on the basis of the reasoning of "working teams of experts" who basically follow certain rules of thumb. Such an approach can be taken as "something is better than nothing." However, we cannot stay at this stage any longer.

The second stage, involving the use of both experts and computers, is at present the most advanced. The participation of an expert is limited to the supplying of proper algorithms in building up the models and the criteria for choosing the best models with optimal complexity. The decisions for contradictory problems are solved according to the multi-objective criteria.

The third stage, "computers without experts," is also called "artificial intelligence systems." The man-machine dialogue system based on the methods of inductive learning algorithms is the most advanced method of prediction and control. It is important that the artificial intelligence systems operate better than the brain by using these computer-aided algorithms. In contrast to the dialogue systems, the decisions in artificial intelligence systems are made on the basis of general requests (criteria) of the human user expressed in a highly abstract metalanguage. The dialogue is transferred to a level at which contradictions between humans are impossible, and, therefore, the decisions are objective and convincing. For example, man can make the requirement that "the environment be clean as possible," "the prediction very accurate," "the dynamic equation most unbiased," and so on. Nobody would object to such general criteria, and man can almost be eliminated from the dialogue of scientific disputes.

In the dialogue systems, the decisions are made at the level of selection of a point in the "Pareto region" where the contradiction occurs. This is solved by using multi-criterion analysis. In artificial intelligence systems, the discrete points of Pareto region are only inputs for dynamic models constructed on the basis of inductive learning algorithms. Ultimately, the computer will become the arbiter who resolves the controversies between users and will play a very important role in simulations.

1.3 Concept of black box

The black box concept is a useful principle of cybernetics. A black box is a system that is too complex to be easily understood. It would not be worthwhile to probe into the nature of interrelations inside the black box to initiate feedback controls. The cybernetic principle of black box, therefore, ignores the internal mechanics of the system but concentrates on the study of the relationship between input and ouput. In other words, the relationship between input and output is used to learn what input changes are needed to achieve a given change in output, thereby finding a method to control the system.

For example, the human being is virtually a black box. His internal mechanism is beyond comprehension. Yet neurologists have achieved considerable success in the treatment of brain disorders on the basis of observations of a patient's responses to stimuli. Typical cybernetic black box control action is clearly discernible in this example. Several complex situations are tackled using the cybernetic principles. Take the case for instance, of predictions of agricultural crop productions. It would involve considerable time and effort to study the various variables and their effect on each other and to apply quantitative techniques of evaluation. Inputs like meteorological conditions, inflow of fertilizers and so on influence crop production. It would be possible to control the scheduling and quantities of various controllable inputs to optimise output. It is helpful to think of the determinants of any "real culture" as it would be the solution of a set of independent simultaneous equations with many unknowns.

Mathematics can be an extremely good tool in exhausting all the possibilities in that it can get a complete solution of the set of equations (or whatever the case may be). Many mathematicians have predicted that entirely new branches of mathematics would someday have to be invented to help solve problems of society - just as a new mathematics was necessary before significant progress could be made in physics. Scientists have been thinking more and more about interactive algorithms to provide the man-machine dialogue, the intuition, the value judgement, and the decision on how to proceed. Computer-aided self-organization algorithms have given us the scope to the present developments and appear to provide the only means for creating even greater cooperative efforts.

2 SELF-ORGANIZATION MODELING

2.1 Neural approach

Rosenblatt [105], [106] gives us the theoretical concept of "perceptron" based on neural functioning. It is known that single-layered networks are simple and are not capable of solving some problems of pattern recognition (for example, XOR problem) [95]. At least two stages are required: *X* —> *H* transformation, and *H* —» *Y* transformation. Although Rosenblatt insists that *X* —> *H* transformation be realized by random links, *H* —> *Y* transformation is more deterministically only realized by learned links where X, H, and *Y* are input, hidden, and output vectors. This corresponds to an *a priori* and conditional probabilistic links in Bayes' formulae:

$$p(y_j) = \sum_{1}^{N} \left[p_0 \prod_{i=1}^{m} p(y_j/x_i) \right]; \quad j = 1, 2, \cdots, n, \tag{1.1}$$

where *po* is an *a priori* link corresponding to the *X* —> *H* transformation, p(yj/xi) are conditional links corresponding to the *H* —> *Y* transformation, *N* is the sample size, *m* and *n* are the number of vector components in *X* and *Y*, respectively. Consequently, the perceptron structures have two types of versions: probabilistic or nonparametric and parametric. Here our concern is parametric network structures. Connection weights among the *H* —> *Y* links are established using some adaptive techniques. Our main emphasis is on an optimum adjustment of the weights in the links to achieve desired output. Eventually neural nets have become multilayered feedforward network structures of information processing as an approach to various problem-solving.

We understand that information is passed on to the layered network through the input layer, and the result of the network's computation is read out at the output layer. The task of the network is to make a set of associations of the input patterns x with the output patterns y. When a new input pattern is put in the configuration, its output pattern must be identifiable by its association.

An important characteristic of any neural network like "adaline" or "backpropagation" is that output of each unit passes through a threshold logic unit (TLU). A standard TLU is a threshold linear function that is used for binary categorization of feature patterns. Nonlinear transfer functions such as sigmoid functions are used as a special case for continuous output. When the output of a unit is activated through the TLU, it mimics a biological neuron turning "on" or "off." A state or summation function is used to compute the capacity of the unit. Each unit is analyzed independently of the others. The next level of interaction comes from mutual connections between the units; the collective phenomenon is considered from loops of the network. Due to such connections, each unit depends on the state of many other units. Such an unbounded network structure can be switched over to a self-organizing mode by using a certain statistical learning law that connects specific forms of acquired change through the synaptic weights, one that connects present to past behavior in an adaptive fashion so positive or negative outcomes of events serve as signals for something else. This law could be a mathematical function - either as an energy function which dissipates energy into the network or an error function which measures the output residual error. A learning method follows a procedure that evaluates this function to make pseudo-random changes in the weight values, retaining those changes that result in improvements to obtain optimum output response. The statistical mechanism helps in evaluating the units until the network performs a desired computation to obtain certain accuracy in response to the input signals. It enables the network to adapt itself to the examples of what it should be doing and to organize information within itself and thereby learn.

Connectionist models

Connectionist models describe input-output processes in terms of activation patterns defined over nodes in a highly interconnected network [24], [107]. The nodes themselves are elementary units that do not directly map onto meaningful concepts. Information is passed through the units and an individual unit typically will play a role in the representation of multiple pieces of knowledge. The representation of knowledge is thus parallel and distributed over multiple units. In a Connectionist model the role of a unit in the processing is defined by the strength of its connections - both excitatory and inhibitory - to other units. In this sense "the knowledge is in the connections," as Connectionist theorists like to put it, rather than in static and monolithic representations of concepts. Learning, viewed within this framework, consists of the revision of connection strengths between units. Back propagation is the technique used in the Connectionist networks - revision of strength parameters on the basis of feedback derived from performance and emergence of higher order structures from more elementary components.

2.2 Inductive approach

Inductive approach is similar to neural approach, but it is bounded in nature. Research on induction has been done extensively in philosophy and psychology. There has been much work published on heuristic problem-solving using this approach. Artificial intelligence is the youngest of the fields concerned with this topic. Though there are controversial discussions on the topic, here the scope of induction is limited to the approach of problem-solving which is almost consistent with the systems theory established by various scientists.

Pioneering work was done by Newell and Simon [96] on the computer simulation of human thinking. They devised a computer program called the General Problem Solver (GPS) to simulate human problem-solving behavior. This applies operators to objects to attain targetted goals; its processes are geared toward the types of goals. A number of similarities and differences among the objective steps taken by computer and subjective ways of a human-operator in solving the problem are shown. Newell and Simon [97] and Simon [113] went on to develop the concepts on rule-based objective systems analysis. They discussed computer programs that not only play games but which also prove theorems in geometry, and proposed the detailed and powerful variable iteration technique for solving test problems by computer.

In recent years, Holland, Holyoak, Nisbett and Thagard [25] considered, on similar grounds, the global view of problem-solving as a process of search through a state space; a problem is defined by an initial state, one or more goal states to be reached, a set of operators that can transform one state into another, and constraints that an acceptable solution must meet. Problem-solving techniques are used for selecting an appropriate sequence of operators that will succeed in transforming the initial state into a goal state through a series of steps. A selection approach is taken on classifying the systems. This is based on an attempt to impose rules of "survival of the fittest" on an ensemble of simple productions.

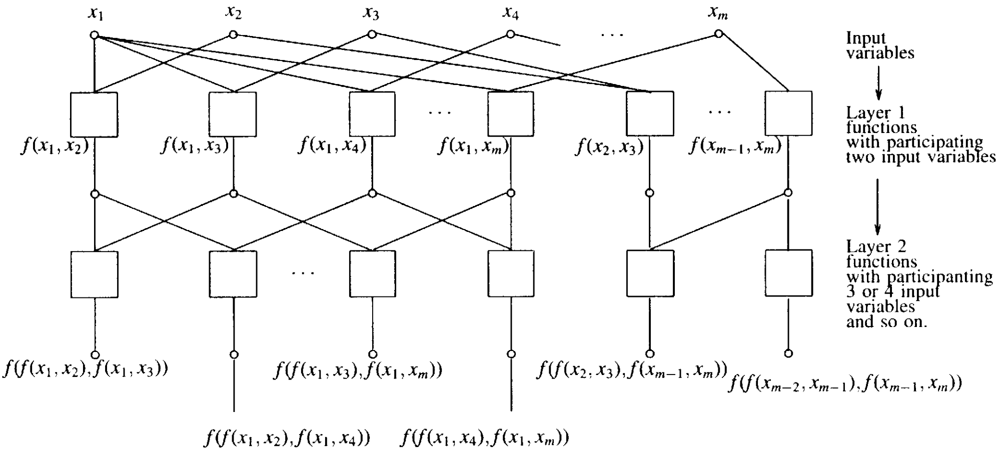
x_1
x_2
x_3
x_4
. . .
x_m
Input variables
Layer 1 functions with participating two input variables
Layer 2 functions with participanting 3 or 4 input variables and so on.
$f(x_1,x_2)$
$f(x_1,x_3)$
$f(x_1,x_4)$
$f(x_1,x_m)$
$f(x_2,x_3)$
$f(x_{m-1},x_m)$
$f(f(x_1,x_2),f(x_1,x_3))$
$f(f(x_1,x_3),f(x_1,x_m))$
$f(f(x_2,x_3),f(x_{m-1},x_m))$
$f(f(x_{m-2},x_{m-1}),f(x_{m-1},x_m))$
$f(f(x_1,x_2),f(x_1,x_4))$
$f(f(x_1,x_4),f(x_1,x_m))$

Figure 1.1. Multilayered induction for gradual increase of complexity in functions

This ensemble is further enhanced by criterion rules which implement processes of genetic cross-over and mutation on the productions in the population. Thus, productions that survive a process of selection are not only applied but also used as "parents" in the synthesis of new productions. Here an "external agent" is required to play a role in laying out the basic architecture of those productions upon which both selective and genetic operations are performed. These classification systems do not require any *a priori* knowledge of the categories to be identified; the knowledge is very much implicit in the structure of the productions; i.e., it is assumed as the *a priori* categorical knowledge is embedded in the classifying systems. The concepts of "natural selection" and "genetic evolutions" are viewed as a possible approach to normal levels of implementation of rules and representations in information processing models.

In systems environment there are dependent (y1,y2,...,y_n) and independent variables (x1,x2,... ,xm). Our task is to know which of the independent variables activate on a particular dependent variable. A sufficient number of general methods are available in mathematical literature. Popular among them is the field of applied regression analysis. However, general methods such as regression analysis are insufficient to account for complex problem-solving skills, but those are backbone for the present day advanced methods. Based on the assumption that composite (control) systems must be based on the use of signals that control the totality of elements of the systems, one can use the principle of induction; this is in the sense that the independent variables are sifted in a random fashion and activated them so that we could ultimately select the best match to the dependent variable.

Figure 1.1 shows a random sifting of formulations that might be related to a specific dependent variable, where f( ) is a mathematical formulation which represents a relationship among them. This sort of induction leads to a gradual increase of complexity and determines the structure of the model of optimal complexity. Figure 1.2 shows another type of induction that gives formulations with all combinations of input variables; in this approach, model of optimal complexity is never missed. Here the problem must be fully defined. The initial state, goal state, and allowable operators (associated with the differences among current state and goal state) must be fully specified. The search takes place step by step at all the units through alternative categorizations of the entities involved in the set up. This type of processing depends on the parallel activity of multiple pieces of emperical knowledge that compete with and complement each other based on an external knowledge in revising

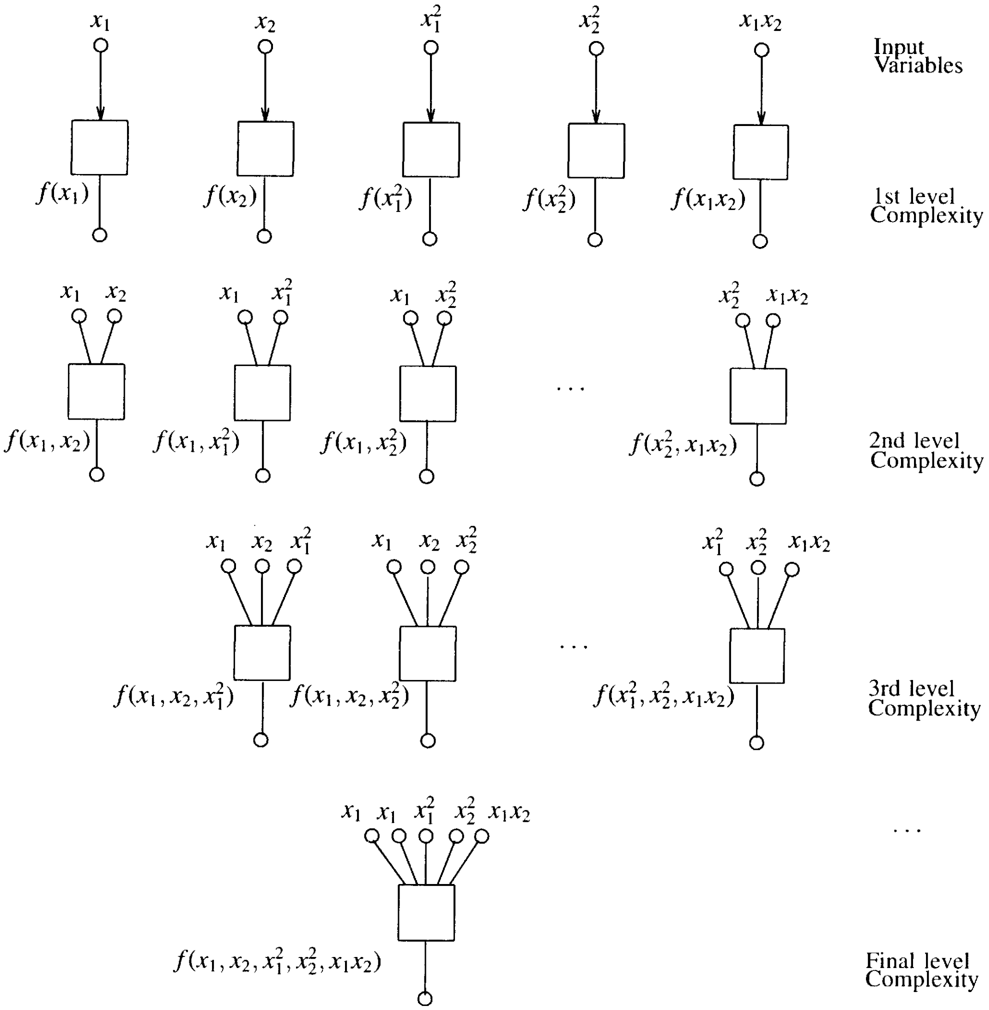

Figure 1.2. Induction of functions for all combinations of input variables

the problem. Such interactive parallelism is a hallmark of the theoretical framework for induction given here.

Simplification of self-organization is regarded as its fundamental problem from the very beginning of its development. The modeling methods created for the last two decades based on the concepts of neural and inductive computing ensure the solution of comprehensive problems of complex systems modeling as applied to cybernetical systems. They constitute an arsenal of means by which—either on the basis of notions concerning system structures and the processes occurring in them, or on the basis of observations of the parameters of these systems—one can construct system models that are accessible for direct analysis and are intended for practical use.

3 INDUCTIVE LEARNING METHODS

Inductive learning methods are also called Group Method of Data Handling (GMDH), Self-organization, sorting out, and heuristic methods. The framework of these methods differs slightly in some important respects. As seen in Chapter 2, the inductive learning algorithms (ILA) have two fundamental processes at their disposal: bounded network connections for generating partial functions and threshold objective functions for establishing competitive learning. The principal result of investigations on inductive learning algorithms (not so much of the examples of computer-designed models presented here), is of a change in view about cybernetics as a science of model construction, in general, and of the role of modern applied mathematics. The deductive approach is based on the analysis of cause-effect relationships. The common opinion is that in the man-machine dialogue, the predominant role is played by the human operator; whereas, the computer has the role of "large calculator." In contrast, in a self-organization algorithm, the role of human operator is passive - he is no longer required to have a profound knowledge of the system under study. He merely gives orders and needs to possess only a minimal amount of *a priori* information such as (i) how to convey to the computer a criterion of model selection that is very general, (ii) how to specify the list of feasible "reference functions" like polynomials or rational functions and harmonic series, and (iii) how to specify the simulation environment; that is, a list of possible variables. The objective character of the models obtained by self-organization is very important for the resolution of many scientific controversies [22]. The man-machine dialogue is raised to the level of a highly abstract language. Man communicates with the machine, not in the difficult language of details, but in a generalized language of integrated signals (selection criteria or objective function). Self-organization restores the belief that a "cybernetic paradise" on earth, governed by a symbiosis between man (the giver of instructions) and machine (an intelligent executer of the instructions) is just around the corner. The self-organization of models can be regarded as a specific algorithm of computer artificial intelligence. Issues like "what features are lacking in traditional techniques" and "how is it compensated in the present theory" are discussed before delving into the basic technique and important features of these methods.

3.1 Principal shortcoming in model development

First of all, let us recollect the important invention of Heisenberg's uncertainty principle from the field of quantum theory which has a direct or indirect influence on later scientific developments. Heisenberg's works became popular between 1925 and 1935 [23], [102]. According to his principle, a simultaneous direct measurement between the coordinate and momentum of a particle with an exactitude surpassing the limits is impossible; furthermore, a similar relationship exists between time and energy. Since his results were published, various scientists have independently worked on Heisenberg's uncertainty principle.

In 1931, Godel published his works on mathematical logic showing that the axiomatic method itself had inherent limitations and that the principal shortcoming was the so-called inappropriate choice of "external complement." According to his well-known incompleteness theorem [126], it is in principle impossible to find a unique model of an object on the basis of empirical data without using an "external complement" [10]. The regularization method used in solving ill-conditioned problems is also based on this theorem. Hence "external complement" and "regularization" are synonyms expressing the same concept.

In regression analysis, the root mean square (RMS) or least square error determined on the basis of all experimental points monotonically decreases when the model complexity gradually increases. This drops to zero when the number of coefficients n of the model becomes equal to the number of empirical points N. Every equation that possesses n coefficients can be regarded as an absolutely accurate model. It is not possible, in principle, to find a unique model in such a situation. Usually experienced modellers use trial and error techniques to find a unique model without stating that they consciously or unconsciously

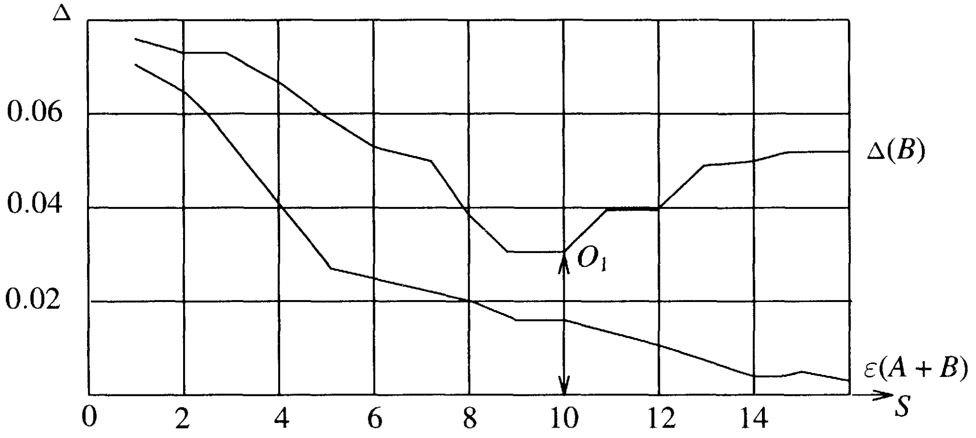

Figure 1.3. Variation in least square error $\varepsilon(A+B)$ and error measure of an "external complement" $\Delta(B)$ for a regression equation of increasing complexity S; O_1 is the model of optimal complexity

use an "external complement," necessary in principle for obtaining a unique model. Hence, none of the investigators appropriately selects the "external complement"—the risk involved in using the trial and error methods.

3.2 Principle of self-organization

In complex systems modeling we cannot use statistical probability distributions, like normal distribution, if we possess only a few empirical points. The important way is to use the inductive approach for sifting various sets of models whose complexity is gradually increased and to test them for their accuracy.

The principle of self-organization can be formulated as follows: When the model complexity gradually increases, certain criteria, which are called selection criteria or objective functions and which have the property of "external complement," pass through a minimum. Achievement of a global minimum indicates the existence of a model of optimum complexity (Figure 1.3).

The notion that there exists a unique model of optimum complexity, determinable by the self-organization principle, forms the basis of the inductive approach. The optimum complexity of the mathematical model of a complex object is found by the minimum of a chosen objective function which possesses properties of external supplementation (by the terminology of Gödel's incompleteness theorem from mathematical logic). The theory of self-organization modeling is based on the methods of complete, incomplete and mathematical induction [4]. This has widened the capabilities of system identification, forecasting, pattern recognition and multicriterial control problems.

3.3 Basic technique

The following are the fundamental steps used in self-organization modeling of inductive algorithms:

1. Data sample of N observations corresponding to the system under study is required; Split them into training set A and testing set B $(N = N_A + N_B)$.

2. Build up a "reference function" as a general relationship between dependent (output) and independent (input) variables.
3. Identify problem objectives like regularization or prediction. Choose the objective rule from the standard selection criteria list which is developed as "external complements."
4. Sort out various partial functions based on the "reference function."
5. Estimate the weights of all partial functions by a parameter estimation technique using the training data set A.
6. Compute quality measures of these functions according to the objective rule chosen using the testing data set *B*.
7. Choose the best measured function as an optimal model. If you are not satisfied, choose *F* number of partial functions which are better than all (this is called "freedom-of-choice") and do further analysis.

Various algorithms differ in how they sift partial functions. They are grouped into two types: single-layer and multi-layer algorithms. Combinatorial is the main single-layer algorithm. Multi-layer algorithm is the layered feedforward algorithm. Harmonic algorithm uses harmonics with nonmultiple frequencies and at each level the output errors are fed forward to the next level. Other algorithms like multilevel algorithm are comprised of objective system analysis and two-level, multiplicative-additive, and multilayer algorithms with error propagations. We go through them in detail in the second chapter. Modified variants of multilayer algorithms were published by Japanese researchers (usually with suggestions regarding their modifications) [78], [122], [108]. Shankar [110] compared the inductive approach with the regression analysis with respect to accuracy of modeling for a small sample of input data. There were other researchers [6], [7], [12], [84], [94], [109] who solved various identification problems using this approach. Farlow [16] compiled various works of US and Japanese researchers in a compendium form. There are a number of investigators who have contributed to the development of the theory and to applications of this self-organization modeling. The mathematical theory of this approach has shown that regression analysis is a particular case of this method; however, comparison of inductive learning algorithms and regression analysis is meaningless.

3.4 Selection criteria or objective functions

Self-organization modeling embraces both the problems of parameter estimation and the selection of model structure. One type of algorithm generates models of different complexities, estimates their coefficients and selects a model of optimal complexity. The global minimum of the selection criterion, reached by inducting all the feasible models, is a measure of model accuracy. If the global minimum is not satisfied, then the model has not been found. This happens in the following cases: (a) the data are too noisy, (b) there are no essential variables among them, (c) the selection criterion is not suitable for the given task of investigation, and (d) time delays are not sufficiently taken into account. In these cases, it is necessary to extend the domain of sifting until we obtain a minimum. Each algorithm uses at least two criteria: an internal criterion for estimating the parameters and an external one for selecting the optimal structure. The external criterion is the quantitative measure of the degree of correspondence of a specific model to some requirement imposed on it. Since the requirements can be different, in modeling one often uses not one but several external criteria; that is, a multicriterion selection. Successive application of the criteria is used primarily in algorithms of objective systems analysis and multilevel long-range forecasting. Furthermore, several criteria are necessary for increasing the noise immunity of the modeling. Selection criteria are also called objective functions or objective rules as they verify and lead to the obtaining of optimal functions according to specified requirements. We can also say that these functions are used to evaluate the threshold capacity of each unit by the quantitative comparison of models of varying complexity necessary for selecting a subset of the best models from the entire set of model candidates generated in the self-organization process. If one imposes the requirement of uniqueness of choice with respect to one or several criteria, then the application of such a criterion or group of criteria yields a unique model of optimal complexity. We give here the typical criteria, historically the first external criteria and their different forms.

Suppose that the entire set (sample) of the original data points N is partitioned into three disjoint subsets A, B and C (parts of the sample) and denotes the union A U B = W. All the criteria used in the algorithms can be expressed in terms of the estimates of the model coefficients obtained on A, B and W and in terms of the estimates of the output variables of the models on A, B, C and W.

We assume that the initial data (N points) are given in the form of matrices below:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_A \\ -- \\ \mathbf{y}_B \\ -- \\ \mathbf{y}_C \end{pmatrix}, \quad X = \begin{pmatrix} X_A \\ -- \\ X_B \\ -- \\ X_C \end{pmatrix}, \quad \begin{array}{c} \mathbf{y}[N \times 1], \\ \\ X[N \times m], \\ \\ N_A + N_B + N_C = N. \end{array} \tag{1.2}$$

The optimal dependence of the output y on the m input variables $\mathbf{x}$ is sought by the group of data handling in inductive fashion the class of functions that are linear in the coefficients of $\mathbf{y} = X\hat{\mathbf{a}}$. The submatrices of matrices X_A and X_B corresponding to any particular model of complexity s (includes s of the m arguments, $s \leq m$), which is tested in the selection process, are of complete rank.

It is convenient to define criteria by some "elementary" quantities. For example, when partitioning a sample into different parts, we introduce the quantities:

$$\hat{\mathbf{a}}_G = (X_G^T X_G)^{-1} X_G^T \mathbf{y}_G; \quad G = A, B, \text{ and } W \tag{1.3}$$

$$\varepsilon_G^2 = \|\mathbf{y}_G - X_G \hat{\mathbf{a}}_G\|^2, \tag{1.4}$$

where ε_G is the least squares error; that is, by the least squares technique the coefficients are estimated using set G; and the error is calculated on the same set.

$$\Delta^2(H) = \Delta^2(H/G) = \|\mathbf{y}_H - X_H \hat{\mathbf{a}}_G\|^2, \tag{1.5}$$

where $H = A, B$; $H \cap G = \emptyset$; the notation $\Delta^2(H/G)$ indicates that the error is calculated on set H of the model—the coefficients of which are estimated on set G.

Regularity criterion

This consists of a squared error calculated on the basis of testing set B.

$$\Delta^2(B) \triangleq \sum_{p \in N_B} (y - \hat{y})_p^2 / \sum_{p \in N_B} y_p^2, \tag{1.6}$$

where $\Delta(B)$ indicates the regularity measure; y and $\hat{y}$ are the desired and estimated outputs, respectively.

Other forms of "regularity criterion" are:

$$\Delta^2(B) \hat{=} \sum_{p \in N_B} (y - \hat{y})_p^2 \tag{1.7}$$

$$\Delta^2(B) \hat{=} \sum_{p \in N_B} (y - \hat{y})_p^2 \Big/ \sum_{p \in N_B} (y_p - \bar{y})^2, \tag{1.8}$$

where $\bar{y}$ is the average value of y. $\Delta^2(B)$ is also renotated as $\Delta^2(B/A)$, which denotes the error of the model calculated on set B using the coefficients obtained on set A. The criterion is given in matrix notation using the Eucledean norm below:

$$\begin{aligned} \Delta^2(B) = \Delta^2(B/A) &= \|\mathbf{y}_B - \hat{\mathbf{y}}_B(A)\|^2 \\ &= (\mathbf{y}_B - X_B\hat{\mathbf{a}}_A)^T\,(\mathbf{y}_B - X_B\hat{\mathbf{a}}_A) \\ &= \|\mathbf{y}_B - X_B\hat{\mathbf{a}}_A\|^2, \end{aligned} \tag{1.9}$$

where $\hat{\mathbf{a}}_A = (X_A^T X_A)^{-1} X_A^T \mathbf{y}_A$, and $\hat{\mathbf{y}}_B(A) = X_B\hat{\mathbf{a}}_A$.

Minimum bias or consistent criterion

This consists of a squared error of difference between the outputs of two models developed on the basis of two distinct sets A and B.

$$\eta_{bs}^2 \hat{=} \sum_{p \in W} (\hat{y}^A - \hat{y}^B)_p^2 \Big/ \sum_{p \in W} y_p^2, \tag{1.10}$$

where $\hat{y}^A$ is the estimated output of the function, obtained on the basis of set A, and $\hat{y}^B$ is the estimated output of the function based on set B. Usually the data with higher values of variance are included into set A, while those with the smaller variance are put in set B. When the model is consistent with exact weights on both sets of the data, then the outputs are equal, $\hat{y}^A = \hat{y}^B$ and $\eta_{bs} = 0$. Therefore, the comparison of the model equations using this criterion $\eta_{bs} \to 0$ enables us to obtain consistent models, it is possible to recover an optimal response which represents a physical law of the system hidden in the noisy experimental data. Similar forms in regularity case can be expressed also as

$$\eta_{bs}^2 \hat{=} \sum_{p \in W} (\hat{y}^A - \hat{y}^B)_p^2 \tag{1.11}$$

$$\eta_{bs}^2 \hat{=} \sum_{p \in W} (\hat{y}^A - \hat{y}^B)_p^2 \Big/ \sum_{p \in W} (y_p - \bar{y})^2 \tag{1.12}$$

$$\begin{aligned} \eta_{bs}^2 &\hat{=} \|\hat{\mathbf{y}}^A - \hat{\mathbf{y}}^B\|_W^2 \\ &= \|X_W\hat{\mathbf{a}}_A - X_W\hat{\mathbf{a}}_B\|^2 \\ &= (\hat{\mathbf{a}}_A - \hat{\mathbf{a}}_B)^T X_W^T X_W (\hat{\mathbf{a}}_A - \hat{\mathbf{a}}_B). \end{aligned} \tag{1.13}$$

Another form of this criterion expresses somewhat a different requirement.

$$\eta_a^2 \hat{=} \|\hat{\mathbf{a}}_A - \hat{\mathbf{a}}_B\|^2, \tag{1.14}$$

where $\hat{a}_A$ and $\hat{a}_B$ are the coefficient vectors estimated on the basis of sets A and B, respectively. The criteria in this group do not take into account the error of the model in explicit form; the criterion of minimum coefficient bias reflects the requirement that the coefficient estimates in the optimal model, calculated on sets A and B, differ only minimally so that they appear to agree. The well-known absolute noise immune criterion is defined as

$$\begin{aligned} V^2 &\hat{=} (X_W\hat{\mathbf{a}}_A - X_W\hat{\mathbf{a}}_W)^T(X_W\hat{\mathbf{a}}_W - X_W\hat{\mathbf{a}}_B) \\ &= (\hat{\mathbf{a}}_A - \hat{\mathbf{a}}_W)^T X_W^T X_W(\hat{\mathbf{a}}_W - \hat{\mathbf{a}}_B) \\ &= (\hat{\mathbf{y}}^A - \hat{\mathbf{y}}^W)_W^T(\hat{\mathbf{y}}^W - \hat{\mathbf{y}}^B)_W, \end{aligned} \tag{1.15}$$

where $\hat{a}_W$ is the coefficient vector obtained on the basis of whole set W. It is possible to select a model that is not sensitive to the data on which it is based. The minimum bias criterion requires the model to yield the same results at successive experimental points of N_A and N_B. It is possible to recover a hidden physical law using this criterion from the noisy data.

Prediction criterion

This consists of a squared error calculated on the basis of a separate examin set C, which is not used in estimating the coefficients:

$$\Delta^2(C/W) \hat{=} \sum_{p \in N_C} (y - \hat{y})_p^2 / \sum_{p \in N_C} y_p^2. \tag{1.16}$$

In case of finite difference equations, the criterion is also evaluated as

$$i^2(W) \hat{=} \sum_{p \in N_W} (y - \hat{y})_p^2 / \sum_{p \in N_W} y_p^2, \tag{1.17}$$

where the estimate $\hat{y}$ is obtained through step by step integration of a difference equation from the given initial conditions. The autoregression form of such model is

$$\hat{y}_p = a_0 + a_1\hat{y}_{p-1} + a_2\hat{y}_{p-2} + \cdots + a_\tau\hat{y}_{p-\tau}. \tag{1.18}$$

This criterion can be also calculated on other parts of the sets $i^2(A)$, and $i^2(B)$.

Combined criteria

Sometimes we choose not only the minimum bias models but also the models with other characteristics. Combined criteria are used as two or more participating criteria in one function. In solving practical problems, it is often necessary to obtain a model that satisfies several requirements simultaneously. Using an objective rule, we then arrive at the familiar problem of multi-criterion selection when some or all of the criteria are contradictory. For example, even the simple problem of selecting a model that will simultaneously be the most regular ("exact") and have the least bias often proves contradictory. Under these conditions, one selects a unique model by using the combined criteria in the form of the sum of the individual criteria with certain weight factors. For example, the combined criteria is given as

$$k^2 \hat{=} \alpha k_1^2 + (1 - \alpha)k_2^2, \tag{1.19}$$

where k_1 and k_2 are given criteria and α is the weight factor. One can also use the normalized values of the criteria:

$$k_l^2 \hat{=} \bar{k}_{1l}^2 + \bar{k}_{2l}^2 = k_{1l}^2/k_{1\,max}^2 + k_{2l}^2/k_{2\,max}^2, \tag{1.20}$$

where l is the index of the model under consideration and the maximum values of the criteria are determined out of all the F models that participate in the sorting.

$$k_{1\,max}^2 = \max_{l \in F} k_{1l}^2, \quad k_{2\,max}^2 = \max_{l \in F} k_{2l}^2. \tag{1.21}$$

The following are some of the combined criteria used in the algorithms.

One of the combined criteria i's "bias plus approximation error." Here the index of the model concerned is not shown.

$$c1^2 \triangleq \bar{\eta}_{bs}^2 + \bar{\varepsilon}^2(W) = \eta_{bs}^2/\eta_{bs\,max}^2 + \varepsilon^2/\varepsilon_{max}^2, \tag{1.22}$$

where $\bar{\varepsilon}^2(W)$ is the normalized approximation error on the data set $W(= A \cup B)$ using the coeficients obtained on W; this is nothing but the least square error. The second form of combined criterion is "bias plus regularity":

$$c2^2 \triangleq \bar{\eta}_{bs}^2 + \bar{\Delta}^2(B). \tag{1.23}$$

Another form of combined criteria which has the best prediction properties ("bias plus error on examination") is

$$c3^2 \triangleq \bar{\eta}_{bs}^2 + \bar{\Delta}^2(C). \tag{1.24}$$

It provides the most unbiased, stable and accurate predicting models, where $\Delta(C) = \Delta(C/W)$ is the mean square error of predictions calculated on data set C using the coefficients obtained on W. The criterion $\Delta(C)$ can also be replaced by $i(W)$; it is appropriate in the case of step-by-step predictions. In calculating criterion $c3$, we can usually divide data in proportions of part A = 40 %, part B = 40 % and part C = 20 %. Sets A and B are used to calculate the minimum bias measure and set C is used for predicting error. In case of criteria $c1$ and $c2$, whole data can be divided into two parts.

There are other forms of combined criteria depending on the combination of various criteria used.

Balance-of-variables criterion

This is used in obtaining a model for long-term predictions in the case of some known *a priori* dependence between variables. For example, if $y = f(x_1, x_2, x_3)$ is the dependent relation, then the balance criterion has the form

$$b \triangleq \sum_{p \in N_C} (y - f(x_1, x_2, x_3))_p^2 / \sum_{p \in N_C} y_p^2, \tag{1.25}$$

where N_c is the set of points in the extrapolation interval and y is the desired output. In the problems, where balance-of-variables is not known, it can be discovered with the help of minimum-of-bias criterion.

Regularity criterion is useful in obtaining an exact approximation of a system as well as of a short-term prediction (for one or two steps ahead) of the processes taking place in it. In the interpolation interval all of the models yield almost the same results (we have the principle of multiplicity of models). In the extrapolation interval the predictions diverge, forming a so called "fan" of predictions.

The minimum-of-bias criterion yields a narrower fan, and hence a longer prediction time than the regularity criterion. This means that prediction is possible for several steps ahead (medium term prediction). However, the theory of self-organization will not solve the problems to which it is applied unless it yielded examples of exact long-term predictions.

The balance-of-variables criterion is proposed for long-range predictions. This requires simultaneous prediction of several interrelated variables. In many examples these variables are constructed artificially. For example, for three variables it is possible to discover the laws:

$$x_1 = f_{11}(x_2, x_3), \quad x_2 = f_{22}(x_1, x_3), \quad x_3 = f_{33}(x_1, x_2), \tag{1.26}$$

where f_{11}, f_{22}, and f_{33} are the functional relations among the variables. The balance-of-variables criterion requires that these relations between pairs of variables be satisfied not only in the interpolation interval, but also in the extrapolation interval. For this purpose, the differences are constructed between "direct" and "inverse" functions. The inverse functions $x_1^* = f_{21}(x_2, x_3)$, $x_2^* = f_{32}(x_1, x_3)$ and $x_3^* = f_{13}(x_1, x_2)$ are computed from the second, third and first laws of the above "direct" equations, respectively. The "inverse" functions can also be obtained as f_{31}, f_{12} and f_{23}. The first subscript is the number of equation and the second is the number of the variable to be determined. If the "direct" and "inverse" functions are exact, the balance criterion requires that

$$b_1 = (f_{11} - f_{21}) \rightarrow 0, \quad b_2 = (f_{22} - f_{32}) \rightarrow 0, \quad b_3 = (f_{33} - f_{13}) \rightarrow 0. \tag{1.27}$$

The balance-of-variables criterion measures the unbalances b_1, b_2, and b_3 in the extrapolation interval as

$$\begin{aligned} B^2 &= \sum_{N_C} [f_{11}(x_2, x_3) - f_{21}(x_2, x_3)]^2 / \sum_{N_C} [f_{11}(x_2, x_3)]^2 \\ &+ \sum_{N_C} [f_{22}(x_1, x_3) - f_{32}(x_1, x_3)]^2 / \sum_{N_C} [f_{22}(x_1, x_3)]^2 \\ &+ \sum_{N_C} [f_{33}(x_1, x_2) - f_{13}(x_1, x_2)]^2 / \sum_{N_C} [f_{33}(x_1, x_2)]^2 \\ &= (b_1^2 + b_2^2 + b_3^2), \end{aligned} \tag{1.28}$$

where NC is number ef points in the prediction or examin data set.

This criterion yields reference points in the future; it requires that a law, effective up to the present, continue into the future in the extrapolation interval; the sum of unbalances in the extrapolation interval should be minimal. In cases where exact relations are not known in the interpolation interval, these can be obtained by using minimum bias criterion in one of the inductive learning algorithms.

The correctness of the prediction is checked according to the values of the criterion. By gradually increasing the prediction time, we arrive at a prediction time for which it is no longer possible to find an appropriate trend in the fan of a given "reference function." The value of the minimum function begins to increase; thus appropriate action must be taken. For example, it may be necessary to change the "reference function." For a richer choice of models, it is also recommended that one go from algebraic to finite-difference equations, take other system variables, estimate the coefficients and others.

3.5 Heuristics used in problem-solving

The term *heuristic* is derived from the Greek word *eureka* (to discover). It is defined as "experiential, judgemental knowledge; the knowledge underlying 'expertise'; rules of thumb, rules of good guessing, that usually achieve desired results but do not guarantee them" [17]. Heuristics does not guarantee results as absolute as conventional algorithms do, but it offers efficient results that are specific and useful most of the time. Heuristic programming provides a variety of ways of capturing human knowledge and achieving the results as per the objectives. There is a slight controversy in using heuristics in building up expert and complex systems studies. Knowledge-base and knowledge-inference mechanisms are developed in expert systems. The performance of an expert system depends on the retrieval of the appropriate information from the knowledge base and its inference mechanism in evaluating its importance for a given problem. In other words, it depends on how effective logic programming and the building up of heuristics is in the mechanisms representing experiential knowledge. The main task of heuristics in self-organization modeling is to build up better man-machine information systems in complex systems analysis thereby reducing man's participation in the decision-making process (with higher degree of generalization.)

Basic modeling problems

Modeling is used for solving the problems: (i) systems analysis of the interactions of variables in a complex object, (ii) structural and parametric identification of an object, (iii) long-range qualitative (fuzzy) or quantitative (detailed) prediction of processes, and (iv) decision-making and planning.

Systems analysis of the interactions of variables precedes identification of an object. It enables us not only to find the set of characteristic variables but also to break it into two subsets: the dependent (output) variables and the independent (input or state) variables (arguments or factors).

In identification, the output variables are given and one will need to find the structure and parameters of all elements. Identification leads to a physical model of the object, and hence can be called the determination of laws governing the object. In the case of noisy data, a physical model can only be used for determining the way the object acts and for making short-range predictions. Quantitative prediction of the distant future using such physical model is impossible. Nevertheless, one is often able to organize a fuzzy qualitative long-range prediction of the overall picture of the future with the aid of so-called loss of scenarios according to the "if-then" scheme. There is a basic difference between the two approaches to modeling. The only way to construct a better mathematical model is to use one's experience ("heuristics or rules of thumb"). Experience, however, can be in the form of the author's combined representations of the model of the object or of the empirical data - the results of an active or passive experiment. The first kind of experiment leads to simulation modeling and the second to the experimental method of inductive learning or self-organization modeling. The classical example of simulation modeling is the familiar model of world dynamics [20]. A weak point with simulation method is the fact that the modeller is compelled to exhibit the laws governing all the elements, including those he is uncertain about or which he thinks are simply less susceptible to simulation. In contrast to simulation modeling, the inductive approach chooses the structure of the model of optimal complexity by testing many candidate models according to an objective function.

In mathematical modeling, certain statistical rules are followed to obtain solutions. These rules, based on certain hypothesis, help us in achieving the solutions. If we take the problem of pattern classification, a discriminant function in the form of a mathematical equation is estimated using some empirical data belonging to two or more classes. The mathematical equation is trained up using a training data set and is selected by one of the statistical criterion, like minimum distance rule. The second part of data of discriminant function is tested for its validation. Here our objective is to obtain optimal weights of the function suited for the best classification; this is mainly based on the criterion used in the procedure, the data used for training and testing the function, and the parameter estimation technique used for this purpose. Obtaining a better function depends on all these factors and how these are handled by an experienced modeller. This depends on the experience and on the building up of these features as heuristics into the algorithm. This shows the role of the human element in the feedback loop of systems analysis.

Developing a mathematical description according to the input-output characteristics of a system, and generating partial functions by linear combinations of the input arguments from the description, splitting of data into number of sets and design of "external complement" as a threshold objective function are noted as common features established in learning mechanism of the inductive algorithms. The output response of the network modeling depends highly on how these features are formed in solving a specific problem. Depending on the researcher's experience and knowledge about the system, these features are treated as heuristics in these algorithms.

Mathematical description of the system

A general relationship between output and input variables is built up in the form of a mathematical description which is an overall form of relationship refering to the complex system under study. This is also called "reference function." Usually the description is considered a discrete form of the Volterra functional series which is also called Kolmogorov-Gabor polynomial:

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk} x_i x_j x_k + \cdots, \tag{1.29}$$

where the output is designated as y, the external input vector as $\mathbf{x} = (x_1, x_2, \cdots)$, and $\mathbf{a}$ the vector of coefficients or weights. This is linear in parameters a and nonlinear in x. Components of the input vector $\mathbf{x}$ could be independent variables, functional terms, or finite difference terms. This means that the function could be either an algebraic equation, a finite difference equation, or an equation with mixed terms. This polynomial represents the full form of mathematical description. This can be replaced with a system of partial polynomials of the form

$$y_i = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2, \tag{1.30}$$

where $i, j = 1, 2, \cdots, m;\ i \neq j$.

Mathematical descriptions can be grouped into three forms as single input-single output forms (trend equations), multi-input-single output forms (multivariate equations), and multi-input-multi-output forms (system of equations). Specific terms like moving averages, logarithmic terms, time function, time, harmonic trends, and so on, can be considered under these descriptions.

(i) When we think about rationalized descriptions according to our understanding of the system, we have to consider interaction of independent variables in the "reference functions." There are various hypotheses regarding the interaction of these variables. For example, there are four variables (x_1, x_2, x_3, x_4).

The first hypothesis is that these variables do not interact with each other; then the description is considered as

$$y = a_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4). \tag{1.31}$$

The second hypothesis is that the first variable x_1 does not interact with the others, but that the others interact among themselves.

$$y = a_0 + f_5(x_1) + f_6(x_2, x_3, x_4). \tag{1.32}$$

The third hypothesis is that the first variable interacts with the second, and the third interacts with the fourth.

$$y = a_0 + f_7(x_1, x_2) + f_8(x_3, x_4);\ \text{etc.}, \tag{1.33}$$

where f_1, f_2, f_3, f_4, and f_5 are first-degree polynomials; f_7 and f_8 are second-degree polynomials; and f_6 is a third-degree polynomial.

The examples are easily continued. Also it is clear from physical considerations that one of the hypotheses is true and that the number of hypotheses is small. Thus, the purpose of optimization of the mathematical form is achieved. Possible combinations of all variables can also be regarded as a sorting of a number of hypotheses—one of which is true. Realization of such combinations cannot lose the optimal model because it ensures complete sorting of all possible models for a given support function.

(ii) One needs to investigate the convergence to trends of the process using multilayered inductive approach. The level of trends can be done using algebraic equations or finite difference analogues of differential equations. The most general systems analysis is based on equations of the form:

$$y^t = f_1(u, t) + f_2(y^{t-1}, y^{t-2}, \cdots, y^{t-\tau}), \tag{1.34}$$

where f_1 is a source function as a linear trend with variables, time t, and a control vartiable u. To simplify the overall investigation, the analysis is broken into two parts:

1. First analysis is of the trends, $y_1^t = f_1(u, t)$; for example, $f_1(u, t) = a_0 + a_1 t + a_2 u$; and
2. Second analysis is of the dynamics, $y_2^t = f_2(y^{t-1}, y^{t-2}, y^{t-3}, \cdots, y^{t-\tau})$.

Although $y^t \neq y_1^t + y_2^t$, to obtain $y^t = y_1^t + y_2^t$ it is necessary to use points of deviations from the first analysis of trends for the second analysis of dynamics.

(iii) If the physical law of the system is considered a "reference function," this would mean that the scope of search for an optimal model in the self-organization modeling is reduced. If there is noise in the empirical data, the physical models cannot guarantee long-range predictions. Our studies show that physical models cannot be used for long-range predictions because of noise in the data. The physical models are suitable only for identification and short-range predictions.

(iv) Sometimes the modeller cannot ascertain which are the output variables in the system. It is very important to find the "leading" variable in the set of output variables of the complex object. The "leading" variable is the variable that is predicted better on-more accurately than the other variables. To identify the "leading" variables certain algorithms are recommended.

(v) Mathematical descriptions with variable coefficients have been used widely as "reference functions" in case of ecological modeling. For example, if we have three control variables $(u_1, u_2$, and $u_3)$, and four other variables $(y_1, y_2, y_3$, and $y_4)$, we can write the complete polynomial as an algebraic equation.

$$\begin{aligned} y_1 = {} & (a_0 + a_1u_1 + a_2u_2 + a_3u_3) + (b_0 + b_1u_1 + b_2u_2 + b_3u_3)y_2 \\ & + (c_0 + c_1u_1 + c_2u_2 + c_3u_3)y_3 + (d_0 + d_1u_1 + d_2u_2 + d_3u_3)y_4. \end{aligned} \tag{1.35}$$

One can include time as a variable along with the control variables in the above form. In the same way, complete polynomials for y_2, y_3, and y_4 can be written. Finite difference form with two delayed arguments is written as

$$\begin{aligned} y_i^{t+1} = {} & (a_0 + a_1u_1 + a_2u_2 + a_3u_3) + (b_0 + b_1u_1 + b_2u_2 + b_3u_3)y_i^t \\ & + (c_0 + c_1u_1 + c_2u_2 + c_3u_3)y_i^{t-1} + (d_0 + d_1u_1 + d_2u_2 + d_3u_3)y_i^{t-2} \end{aligned} \tag{1.36}$$

for i = 1,2, 3,4. These types of polynomials are also used in studies of inflation stability.

(vi) One must take necessary care when the mathematical description is described. The following are four features to improve, in a decisive manner, the existing models of complex objects and to give them an objective character.

1. Descriptions that are limited to a certain class of equations and to a certain form of support functions lead to poor informative models with respect to their performance on predictions. For example, a difference equation with a single delayed argument with constant coefficients is considered a "reference function":

$$x_i^t = a_0 + a_1 x_i^{t-1} + a_2 t + a_3 t^2. \tag{137}$$

   The continuous analogous of such equation is first-order differential equation; the solution of such equation is an exponential function. If many variants are included in the description, the algorithm sorts out the class of equations and support functions according to the choice criteria.
2. If the descriptions are designed with arbitrary output or dependent variables, then output variables are unknown. Those types of descriptions lead to biased equations. Inductive learning algorithms with special features are used to choose the leading variables.
3. There is a wrong notion that physical models are better for long-range predictions. The third feature of the algorithms is that nonphysical models are better for long-range predictions of complex systems. Physical models (that is, models isomorphic to an object which carry over the mechanism of its action), in the case of inexact data are unsuitable for quantitative long-range prediction.
4. The variables which hinder the object of the problem must be recognized. The fourth feature of the algorithms is that predictions of all variables of interest are found as functions of "leading" variables.

Splitting data into training and testing sets

Most of the selection criteria require the division of the data into two or more sets. In inductive learning algorithms, it is important to efficiently partition the data into parts (the efficiency of the selection criteria depends to a large extent on this). This is called "purposeful regularization." Various ways of "purposeful regularization" are as below:

1. The data points are grouped into a training and a checking sequence. The last point of the data belongs to the checking sequence.
2. The data point are grouped into training and checking sequences. The last point belongs to the training sequence.
3. The data points are arranged according to the variance and are grouped into training and testing parts. This is the usual method of splitting data. Half of the data with the higher values is used as the training set and another half is used as the testing set.
4. The data points represent the last year. Points correspond to the past data for all years that differ from the last by a multiple of prediction interval T_{pre}. For example, the last year in the data table corresponds to the year 1990; prediction interval is made for the year 1994 (ie., T_{pre} = 4 years). The checking sequence comprises the data for the years 1990, 1986, 1982, 1978, etc. and the other data belong to the training sequence.

5. The checking sequence consists of only one data point. For example, if we have data of N years and the prediction interval is T_{pre}, then the points from 1 to $N — T_{pre} — 1$ belong to the training sequence and Nth point belongs to the checking sequence. This is used in the algorithm for the first prediction.

   The second prediction is obtained based on the same algorithm, with another checking point which consists of $N — 1$ point; the training sequence contains from 1 *to* $N — T_{pre} — 2$ points.

   The third prediction is based on the (N- 2)nd point for checking sequence and 1 *to* $N — T_{pre} — 3$ points for training sequence.

   The predictions are repeated ten to twenty times and one obtains prediction polynomials. All the polynomials are summed up and taken average of it. Each prediction is made for an interval length of T_{pre}, and the series of prediction equations is averaged.
6. The data points are grouped into two sequences: the last points in time form the training sequence; and the checking sequence is moved backward l years, where l/ depends on the prediction time and on the number of years for which the prediction is calculated; i.e., it indicates the length of the checking sequence.

Although each method has its unique characteristics of obtaining the model in optimal complexity, only under special conditions are they used. The most usual method is the third method which has to do with the variance and helps minimize the selection layers in case of multi-layer inductive approach.

The following are some examples to show the effect of partitioning of data.

1. It is the method of optimization of allocation of data sample to training and testing sets. There were 14 points in the data sample. Experiments were conducted with different proportions of training and testing sets to obtain the optimal model using the regularity criterion. Figure 1.4 illustrates that a choice of proportionality 9:5 is optimal from the point of view of the number of selection layers in the multilayer iterative algorithm. The simplest and most adequate model was obtained with such an allocation of points. It was noted that the regularity criterion could be taken as the reciprocal of the mean square error in the testing set.
2. Here is another example of the effect of partitions on the global minimum achieved by using the combined criterion c3 that is defined as

$$c3^2 \triangleq \alpha\, \eta_{bs}^2 + (1-\alpha)\,\Delta^2(C), \tag{1.38}$$

   where

$$\eta_{bs}^2 \triangleq \sum_{N} (\hat{y}^A - \hat{y}^B)^2 \Big/ \sum_{N} y^2,$$
$$\Delta^2(C/W) \triangleq \sum_{N_C} (y - \hat{y})^2 \Big/ \sum_{N_C} y^2.$$

   A random data of 100 points is arranged as per its variance and is divided into proportions $A : B : C$, as shown in the Table 1.3. The combined criterion measure at each layer is given for different values of α. Global minimum for each experiment is indicated with "*". When $\alpha = 1$, only minimum bias criterion is participated. As the value of a decreases, the participation of $\Delta^2(C)$ increases in selecting the optimal model. From the global values of the criteria, one can note that the optimum splitting of data is 45:45:10.
3. One of the experiments was done by finding the required partition of empirical data points using the extremal values of the minimum bias selection criterion on the set of

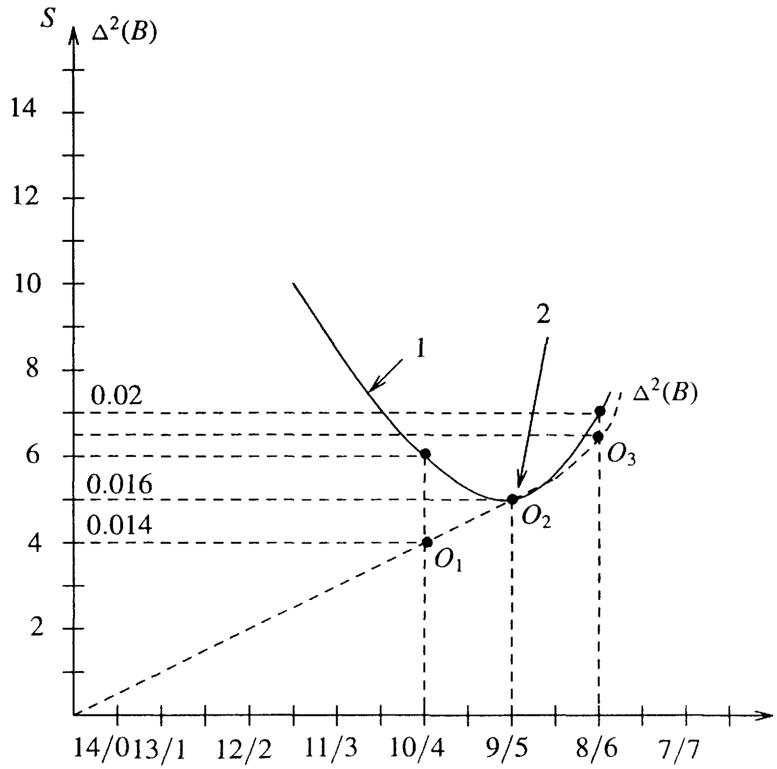

Figure 1.4. Optimum allocation of data to training and testing sets, where S is the number of selection layers, $\Delta^2(B)$is the error measure using regularity criterion. 1. plot of number of selection layers and 2. chosen optimum allocation

all possible versions of data partition in a prescribed relationship [128]. It was shown that the different possible partitions effect the global minimum.

Objective functions

Thinking of objectives in mathematical form is one of the difficult tasks in these algorithms. Extensive has been work done in this direction and enormous contributions have been made to the field in recent years. Most of the objective functions are related to the standard mathematical modeling objectives such as regularization, prediction, unbiasedness and so on. There are standard statistical criteria used by various researchers according to statistical importance, One can also design his own set of criteria with regard to specific objectives. The following is a brief sketch of the development of these functions.

(i) In the beginning stages of self-organization modeling (1968 to 1971), it was applied to pattern recognition, identification, and short-range prediction problems. These problems were solved by regularity criterion only.

$$\Delta^2(B) \triangleq \sum_{p\in N_B} (y-\hat{y})^2_p / \sum_{p\in N_B} y^2_p, \tag{1.39}$$

Table 1.3. c3 values for different values of *a* with different partitions

A:B:C	Layer: 1	2	3	4	5	6	7
$\alpha = 1$:							
45:45:10	0.152	0.053	0.073	0.007*	0.120	0.048	0.034
40:40:20	0.176	0.052*	0.146	0.099	0.126	0.158	0.149
35:35:30	0.181	0.151	0.109	0.059*	0.193	0.097	0.159
$a = 0.75$:							
45:45:10	0.323	0.262*	0.360	0.362	0.440	0.440	0.439
40:40:20	0.293	0.249	0.233*	0.306	0.242	0.263	0.265
35:35:30	0.307	0.300	0.313	0.281*	0.452	0.374	0.368
$a = 0.5$:							
45:45:10	0.416	0.423	0.390	0.332*	0.409	0.400	0.373
40:40:20	0.376	0.351	0.346*	0.389	0.407	0.408	0.462
35:35:30	0.389	0.347*	0.362	0.405	0.370	0.370	0.359
$a = 0.25$:							
45:45:10	0.489	0.420	0.384	0.385	0.335*	0.369	0.436
40:40:20	0.443	0.423	0.380*	0.469	0.468	0.467	0.428
35:35:30	0.455	0.427	0.420	0.417*	0.471	0.427	0.453

where y is the desired output variable, y is the estimated output based on the model obtained on training set A (about 70% of data), and N_B, is the number of points in the testing set (about 30% of data) used for computing regularity error.

Sometimes this criterion was used in the form of a correlation coefficient between y and y variables or in the form of a correlation index (for nonlinear models).

(ii) Later, during 1972 to 1975, the ideas of multicriteria choice of models were developed in pattern recognition theory, minimum bias, balance of variables, and combined criteria. Minimum bias criterion is recommended to obtain a physical model; balance-of-variables criterion is preferred to identify a model for long-range predictions. Various criteria like prediction criterion and criteria for probabilistic stability were also proposed during this period. We were convinced that the wide use of the minimum bias and balance of variables criteria, together with the solution of the noise resistance problem, were the major ways of improving the quality of the models.

(iii) During the eighties, there was fruitful research in the direction of developing noise immune criteria which lead to the successful development of various algorithms such as objective system analysis and multilevel algorithms. The noise stability of self-organization modeling algorithms and noise immune external criteria will be discussed in Chapter 3.

There is confusion with the notations used for the selection criteria as developments progressed through the years. Here we try to give various forms of criteria with standard notations.

All the individual criteria, which are of quadratic form, are divided into two basic groups:

(i) accuracy criteria, which express the error in the model being tested on various parts of the sample (example, regularity),

(ii) matching (consistent) criteria, which are a measure of the closeness of the estimates obtained on different parts of the sample (example, minimum bias).

By adding other two groups, such as balance and dynamics (step-by-step integral) criteria, all external criteria are classified into four groups, as given in the Table 1.4, where 3 is the parameter used in averaging the term and $\hat{\mathbf{y}}_W(y_0, \hat{a}_W)$ is the step-by-step integrated

Table 1.4. External criteria

Criteria group	Group symbol	Criteria	Notation Old	Notation New	Computational formula
Accuracy criteria	A	(i) Regularity	$\Delta^2(B)$	AB	$\|\mathbf{y}_B - X_B\hat{\mathbf{a}}_A\|^2$
		(ii) Symmetric (dual) regularity	d^2	AD	$\|\mathbf{y}_A - X_A\hat{\mathbf{a}}_B\|^2 +$ $\|\mathbf{y}_B - X_B\hat{\mathbf{a}}_A\|^2$
		(iii) Stability	S^2	AS	$\|\mathbf{y}_W - X_W\hat{\mathbf{a}}_A\|^2 +$ $\|\mathbf{y}_W - X_W\hat{\mathbf{a}}_B\|^2$
Balance criteria	B	(i) Predictions balance (linear)	B^2	BL	$\|\hat{Q} - (\hat{q}_1 \| \cdots \| \hat{q}_L)\beta\|^2$
		(ii) Variables balance	B^2	BV	$\|\hat{\Phi} - \psi(\hat{y})\|^2$
Matching criteria	C	(i) Consistency (mini mum- bias)	η^2_{bs}	CB	$\|X_W\hat{\mathbf{a}}_A - X_W\hat{\mathbf{a}}_B\|^2$
		(ii) Unbiasedness in coefficients	η^2_a	CC	$\|\hat{\mathbf{a}}_A - \hat{\mathbf{a}}_B\|^2$
		(iii) Variability (absolute noise immune)	V^2	CV	$(X_W\hat{\mathbf{a}}_A - X_W\hat{\mathbf{a}}_W)^T$ $(X_W\hat{\mathbf{a}}_W - X_W\hat{\mathbf{a}}_B)$
Dynamic criteria	D	(i) Integral	i^2	DI	$\|\mathbf{y}_W - \hat{\mathbf{y}}_W(y_0, \hat{\mathbf{a}}_W)\|^2$

Table 1.5. Classification of criteria

No.	Internal	External
i	*Accuracy type*: (a) Mean square error $\varepsilon^2 = \sum_{i \in N_W} (\hat{y} - y)^2_i$ (b) Correlational $\rho_W = \| \text{Cov } (y\hat{y})\|_1^{N_W}$ (c) Distance $d = \sigma(x_i, x_j)$	(a) Ideal $J = \sum_{i \in N_W} (\hat{y} - \overset{\circ}{y})^2$ (b) Correlational $\rho_B = \| \text{Cov } (y\hat{y})\|_1^{N_B}$ (c) Regularity $\Delta^2(B) = \sum_{i \in N_B} (y - \hat{y}^A)^2_i$
ii	*Integral type(Dynamic)*: Stepwise prediction $i^2(W) = \sum_{t=1}^{N_W - 1} (\hat{y}_{t+1} - ay_t)^2$	Stepwise prediction $i^2(C) = \sum_{t=1}^{N_C} (\hat{y}_{t+1} - a\hat{y}_t)^2$
iii	*Differential type (Balance and Matching)*: - -	(a) Balance-of-predictions $B^2 = \sum_{i \in N} [\hat{Q} - \frac{1}{4}(\hat{q}_1 + \hat{q}_2 + \hat{q}_3 + \hat{q}_4)]^2_i$ (b) Minimum- bias or consistency $\eta_{bs} = \sum_{i \in N_W} (\hat{y}^A - \hat{y}^B)^2_i$

output value which is initialized with the first value *yo* using the estimated coefficients *aw*. "Symmetric" and "nonsymmetric" forms of certain criteria are shown. "Symmetric" criterion means one in which the data information in parts A and *B* of the sample are used equally; when it is not, the criterion is "nonsymmetric." These are further discussed in later chapters. Here we have given old and new notations of these criteria; the old notation is followed throughout the book. The new notation will be helpful in following the literature

As it is clear that the internal criteria are the criteria that participate in the interpolation region in estimating or evaluating the parameters of the models; on the other hand, the external criteria are the criteria that use the information from the extrapolation region (partially or fully) in evaluating the models. Table 1.5 demonstrates some of these criteria, where $\overset{o}{y}$ is the ideal output value (without noise).

The inductive approach proposes a more satisfactory way to find optimum decisions in self-organization models for identification and for short- and long-range predictions. This is particularly useful with noisy data. Communication theory and inductive theory differ from one another by the number of dimensions used in self-organization modeling, but they have common analogy according to the principle of self-organization. The internal criteria currently used in the traditional theories does not allow one to distinguish the model of optimal complexity from the more complex overfitted ones.