Estimation of Dependences Based on Empirical Data

Second Edition

Vladimir Vapnik

Information Science & Statistics

Information Science and Statistics

Information Science and Statistics

Akaike and Kitagawa: The Practice of Time Series Analysis.
Cowell, Dawid, Lauritzen, and Spiegelhalter: Probabilistic Networks and Expert Systems.
Doucet, de Freitas, and Gordon: Sequential Monte Carlo Methods in Practice.
Fine: Feedforward Neural Network Methodology.
Hawkins and Olwell: Cumulative Sum Charts and Charting for Quality Improvement.
Jensen: Bayesian Networks and Decision Graphs.
Marchette: Computer Intrusion Detection and Network Monitoring: A Statistical Viewpoint.
Rubinstein and Kroese: The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation, and Machine Learning.
Studený: Probabilistic Conditional Independence Structures.
Vapnik: The Nature of Statistical Learning Theory, Second Edition.
Wallace: Statistical and Inductive Inference by Minimum Massage Length.

Vladimir Vapnik

Estimation of Dependences Based on Empirical Data

Reprint of 1982 Edition

Empirical Inference Science

Afterword of 2006

Springer

Vladimir Vapnik
NEC Labs America
4 Independence Way
Princeton, NJ 08540
vlad@nec-labs.com

Samuel Kotz (*Translator*)
Department of Engineering Management
and Systems Engineering
The George Washington University
Washington, D.C. 20052

Library of Congress Control Number: 2005938355

ISBN 978-1-4419-2158-1 ISBN 978-0-387-34239-9 (eBook)
DOI 10.1007/978-0-387-34239-9

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Vladimir Vapnik

Estimation of Dependences Based on Empirical Data

Translated by Samuel Kotz

With 22 illustrations

Springer

To my daughter

Preface

Estimating dependences on the basis of empirical data has been, and will probably remain, a central problem in applied analysis. This problem is a mathematical interpretation of one of the basic questions of science: how to extract the existing law-like relationship from scattered data.

The simplest attack on this problem is to construct (estimate) a function from its values at certain points. Here we will formulate some general principles of estimating a functional dependence, and then develop an algorithm for the estimation using these principles.

Usually, when one seeks a general principle, intended for a solution of a wide class of problems, one focuses first upon the simplest, most basic problem. This simple version of the problem is treated theoretically with great thoroughness and the scheme obtained for a solution is then extended to all the problems of the class under consideration.

When studying the estimation of functional dependences, the functions which take only one value (i.e., constants) are usually chosen as the simplest problem. One assumes that the measurements of a constant are subject to errors. Given several such measurements, one must determine this constant. There are various ways to state this problem specifically. These are based on different models of measurements with errors. However, regardless of the model, the study of the basic problem leads to the following classical principle of estimating functional dependence based on empirical data:

> Select, from an admissible set of functions, the one which yields the best approximation of the totality of the available empirical data.

This principle is sufficiently general. It leaves the measure of the quality of the relation between the function and the empirical data undefined. Various definitions of this measure are available; for example, the amount of the mean

square deviation of the functional values, the amount of the mean deviation, the maximal deviation, etc. Each definition generates its own method of estimating dependences, such as the least-squares method, the least absolute values method, etc. However, in all cases the principle of the solution (i.e., the search for a function which best approximates the data) remains unchanged.

The main content of this book deals with a study of a different, nonclassical principle of estimating dependences:

> Select, from an admissible set of functions, a function which fulfills a definite relationship between a quantity characterizing the quality of the approximation and a quantity characterizing the "complexity" of the approximating function.

This principle may need some clarification. With increasing complexity of the approximating function, one obtains successively better approximations to the available data, and may even be able to construct a function which will pass through all of the given points. This new principle, unlike the classical one, asserts that we should not strive to get close to empirical data at all costs; that is, we should not excessively complicate the approximating function. For any given amount of data, there exists a specific relationship between the complexity of the approximating function and the quality of the approximation thus obtained. By preserving this relationship, the estimated dependence most accurately characterizes the actual (unknown) dependence. Further improvements of the approximation by increasing the complexity may result in the estimated function approximating the given data better, but representing the actual function less accurately. This nonclassical principle of estimation reflects an attempt to take into account that dependence is estimated with a limited amount of data.

The idea that, with a limited amount of data, the selected function should not merely approximate empirical data but also possess some extremal properties has existed for a long time. It first received theoretical justification in the investigation of the problems of pattern recognition. The mathematical statement of pattern recognition necessarily leads to estimating a function which admits not one (as is the case in our basic problem) but two values. This additional complexity is unexpectedly of fundamental importance. The set of functions taking on two values is much more "varied" than the set of constants (i.e., functions taking on one value).

The important point is that the structure of the set of constant functions is "simple and homogeneous", while that of the set of functions taking on two values is rich and admits ordering according to its complexity. The latter is essential for estimating dependences with limited amounts of empirical data.

Thus the study of pattern recognition problems has shown that the simplest classical problem does not encompass all the problems of estimating dependences, since the class of functions associated with estimating a constant is so limited that no problem of its stratification arises.

The simplest problem of this book is the problem of pattern recognition. We use methods based on classical ideas of statistical analysis as well as those associated with the nonclassical principle of estimation for its solution. All of these methods are adopted for two other problems of estimation: regression estimation and interpretation of the results of indirect experiments.

For our new basic problem, we distinguish between two formulations: estimating functions and estimating values of a function at given points. (These two formulations coincide in the case of estimation of constants.) We distinguish between these formulations since, with a limited amount of data, there may not be enough information to estimate a function satisfactorily as a whole, but at the same time it may be enough to estimate k numbers—the values of a function at given points.

Thus this book is devoted to problems of estimating dependences with limited amounts of data. The basic idea is as follows: the attempt to take into account the fact that the amount of empirical data is limited leads us to the nonclassical principle of estimating dependences. Utilizing this principle allows us to solve delicate problems of estimation. These include determination of optimal set of features in the case of pattern recognition, determination of the structure of the approximating function in the case of regression estimation, and construction of regularizing functions for solving ill-posed problems of interpretation of indirect experiments (i.e., problems which arise due to the limited amount of data and which cannot be solved within the framework of classical setups).

The book contains ten chapters. Chapters 1 and 2 are introductory. In these, various problems of estimating dependences are considered from the common positions of minimizing the expected risk based on the empirical data and various possible approaches to minimizing risks are discussed.

Chapters 3, 4, and 5 are devoted to the study of classical ideas of risk minimization: estimating probability density functions by means of parametric methods and utilization of this density for minimization of the risk. Chapter 3 applies these ideas to pattern recognition problems. Chapters 4 and 5 apply them to regression estimation problems. Beginning with Chapter 6 nonclassical methods of minimization of risk are studied. Chapters 6 and 7 establish the conditions for applying the method of minimization of empirical risk to solutions of problems of minimization of the expected risk for samples of limited size, while Chapters 8–10 utilize these conditions to construct a method of risk minimization based on limited data: the so-called method of structural minimization. (In Chapter 8, we consider the application of the method of structural risk minimization to the problems of pattern recognition and regression. In Chapter 9, we give an application to the solutions to ill-posed problems of interpreting results of indirect experiments. In Chapter 10, we investigate the problem of estimating values of functions at given points based on structural minimization). Finally, Addenda I and II are devoted to algorithms for structural risk minimization.

This book is intended for a wide class of readers: students in upper-level

courses, graduate students, engineers, and scientists. The exposition is such that the proofs do not interfere with the basic flow of the arguments. However, all of the main assertions are proved *in toto*.

We try to avoid generalizations which are possibly important but less indicative of the basic ideas developed in this book. Therefore, in the main part of the book we consider only simple cases (such as quadratic loss functions, equally spaced observations, independent errors, etc.). As a rule, the corresponding generalizations may be achieved using standard methods. The most important of these generalizations concerning arbitrary loss functions are given at the end of the respective chapters.

The main part of the book does not require a knowledge of special branches of mathematics. However, in order to follow the proofs the reader should possess some experience in dealing with mathematical concepts.

The book is not a survey of the standard theory, and it may be biased to some extent. Nevertheless, it is our hope that the reader will find it interesting and useful.

Moscow, 1982 V. VAPNIK

Contents

Chapter 1

The Problem of Estimating Dependences from Empirical Data

§1 The Problem of Minimizing the Expected Risk on the Basis of Empirical Data

Each time a problem of selecting a functional dependence arises the same model is considered: among the totality of possible dependences it is necessary to find one which satisfies a given quality criterion in the best possible manner. Formally this means that on a vector space Z a class of functions $\{g(z)\}$, $z \in Z$, (the class of possible dependences) is given, and a functional

$$I = I(g) \tag{1.1}$$

is defined which is the criterion of quality of the chosen dependence. It is then required to find $g^*(z)$ belonging to $\{g(z)\}$ such that it will minimize the functional (1.1). (We shall assume that the minimum of the functional corresponds to the best quality and that the minimum of (1.1) exists in $\{g(z)\}$.) In the case when the class of functions $\{g(z)\}$ and functional $I(g)$ are explicitly given, the search for $g^*(z)$ which minimizes $I(g)$ is the subject of the calculus of variations.

In this book another case is considered, namely when a probability density function $P(z)$ is defined on Z and the functional is defined as the mathematical expectation †

$$I(g) = \int \Phi(z, g(z))P(z)\,dz. \tag{1.2}$$

† For the sake of simplicity we require the existence of a density. For the main part of the theory to be valid, the existence of a probability measure is sufficient.

The problem is to minimize the functional (1.2) in the case when $P(z)$ is unknown but when a sample

$$z_1, \ldots, z_l \tag{1.3}$$

of observations resulting from random and independent trials according to $P(z)$ is available.

Below in Sections 2, 3, and 4 we shall verify that all the basic problems in estimating functional dependences are reduced to a minimization of (1.2) based on empirical data (1.3). Meanwhile we shall note that there is a substantial difference between problems arising when the functional (1.1) is minimized and those encountered when the functional (1.2) is minimized, on the basis of empirical data (1.3). In the case of minimizing (1.1) the problem is to organize the search for a function $g^*(z)$ belonging to the class $\{g(z)\}$ which minimizes (1.1). When (1.2) is minimized on the basis of the data (1.3), the basic problem is to formulate a constructive criterion for choosing the function rather than organizing a search of the function in $\{g(z)\}$. (The functional (1.2) by itself cannot serve as a criterion for choosing, since the density $P(z)$ appearing in it is unknown.) Thus in the first case the question is "How do we obtain the minimum of a functional in a given class of functions?" while in the second the question is "What should be minimized in order to select from $\{g(z)\}$ a function which will assure that the functional (1.2) will be 'small'?"

The minimization of the functional (1.2) on the basis of the data (1.3) is a problem of mathematical statistics. We shall call it *the problem of minimizing the expected risk on the basis of empirical data.*

When formulating the minimization problem for the expected risk, the class of functions $\{g(z)\}$ will be given in the parametric form $\{g(z, \alpha)\}$.† Here α is a parameter belonging to the set Λ whose specific value $\alpha = \alpha^*$ defines a specific function $g(z, \alpha^*)$ belonging to the class $g(z, \alpha)$. To find the required function means to determine the required value of the parameter α. The study of only a parametric class of functions is not a serious restriction on the problem, since the set Λ to which the parameter α belongs is arbitrary: it can be a set of scalar quantities, of vectors, or of abstract elements.

In terms of the new notation the functional (1.2) is rewritten as

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz, \qquad \alpha \in \Lambda, \tag{1.4}$$

where

$$Q(z, \alpha) = \Phi(z, g(z, \alpha)).$$

The function $Q(z, \alpha)$—which depends on two groups of variables z and α—is called the *loss function.*

† Below we shall always omit the braces when writing a class of functions. A single function is distinguished from a class of functions by indicating whether the parameter α is fixed or not.

The problem of minimizing the expected risk admits a simple interpretation: it is assumed that each function $Q(z, \alpha^*)$, $\alpha^* \in \Lambda$ (i.e., each function of z for a fixed $\alpha = \alpha^*$) determines the amount of the loss resulting from the realization of vector z. The expected loss (with respect to z) for the function $Q(z, \alpha^*)$ is thus determined by the integral

$$I(\alpha^*) = \int Q(z, \alpha^*) P(z)\, dz.$$

The problem is to choose in $Q(z, \alpha)$ a function $Q(z, \alpha^*)$ which minimizes the expected loss when random independent observations $z_1, \ldots, z_l$ from an unknown probability distribution of z are given.

This problem is rather general. We shall now state a particular case. In this case the vector z consists of $n + 1$ coordinates, the coordinate y and n coordinates $x^1, \ldots, x^n$ which form the vector x. The loss function $Q(z, \alpha)$ is given in the form

$$Q(z, \alpha) = \Phi(y - F(x, \alpha)),$$

where $F(x, \alpha)$ is a parametric class of functions. It is necessary to minimize the functional

$$I(\alpha) = \int \Phi(y - F(x, \alpha)) P(x, y)\, dx\, dy, \tag{1.5}$$

when the density $P(x, y)$ is unknown but a random independent sample of pairs

$$x_1, y_1; \ldots; x_l, y_l \tag{1.6}$$

(*the training sequence*) is given.

The problem of minimizing the functional (1.5) on the basis of the empirical data (1.6) is called *the problem of estimating a functional dependence*, and is the subject of this book.† Three basic problems of estimating functional dependences are considered:

(1) the problem of pattern recognition,
(2) the problem of regression estimation,
(3) the problem of interpreting results obtained from indirect experiments.

In the succeeding sections we shall verify that all these problems are reduced to a minimization of the functional (1.5) on the basis of the empirical data (1.6).

† Below we shall use a quadratic loss function $\Phi(y - F(x, \alpha)) = (y - F(x, \alpha))^2$. However, the basic results to be obtained herein do not depend upon the form of loss function.

§2 The Problem of Pattern Recognition

The problem of pattern recognition was formulated in the late 1950s. In essence it can be stated as follows: a person (the instructor) observes occurring situations and determines to which of k classes each one of them belongs. It is required to construct a device which, after observing the instructor's procedure, will carry out the classification approximately in the same manner as the instructor.

Using formal language this statement can be expressed simply as follows: in a certain environment which is characterized by a probability density function $P(x)$, situations x appear randomly and independently. The instructor classifies these situations into one of the k classes. (For simplicity we shall assume in what follows that $k = 2$; this assumption does not limit the generality, since by subsequent subdivisions of situations into two classes one can obtain a subdivision into k classes as well.) Assume that the instructor carries out this classification using the conditional probability distribution function $P(\omega|x)$, where $\omega = \{0, 1\}$ ($\omega = 0$ indicates that the instructor assigns situation x to the first class, and $\omega = 1$ that he assigns it to the second class). Neither the properties of the environment $P(x)$ nor the decision rule $P(\omega|x)$ is known. However, it is known that both functions exist.

Now let a parametric set of functional dependences $F(x, \alpha)$ (the class of decision rules) be given. All functions in the class $F(x, \alpha)$ are indicator functions, i.e., they take on only the two values zero or one. By observing l pairs

$$x_1, \omega_1; \ldots; x_l, \omega_l$$

(the situation being x, and instructor's reaction ω), it is required to choose in the class of indicator functions $F(x, \alpha)$ a function for which the probability of classification different from the instructor's classification is minimal. In other words, the minimum of the functional

$$I(\alpha) = \sum_{\omega = 0, 1} \int (\omega - F(x, \alpha))^2 P(\omega|x) P(x)\, dx$$

must be attained. The functional $I(\alpha)$ will be written in the form

$$I(\alpha) = \int_{X, \omega} (\omega - F(x, \alpha))^2 P(x, \omega)\, dx\, d\omega$$

and the function $P(x, \omega) = P(\omega|x)P(x)$ will be called the joint density of the pair x, ω defined on the space X, ω.

The problem of pattern recognition has thus been reduced to the problem of minimizing the expected risk on the basis of empirical data. The special

feature of this problem is that the class of functions $Q(z, \alpha)$ is not as arbitrary as in the general case. The following restrictions are imposed:

(1) The vector z consists of $n + 1$ coordinates: coordinate ω, which takes on only two values (zero and one), and n coordinates $x^1, \ldots, x^n$ which form the vector x.
(2) The class of functions $Q(z, \alpha)$ is given by

$$Q(z, \alpha) = (\omega - F(x, \alpha))^2,$$

where $F(x, \alpha)$ also takes on only the two values zero and one.

Thus in the pattern recognition problem the value of the loss function is either zero or one. This particular feature of the risk minimization problem characterizes the pattern recognition problem.†

§3 The Regression Estimation Problem

Two sets of elements X and Y are connected by a functional dependence if to each element $x \in X$ there corresponds uniquely an element $y \in Y$. This relationship is called a function if the set X is a set of vectors and the set Y is that of scalars. However, there exist relationships (dependences) where to each vector x there corresponds a number y which is obtained as a result of random trials according to the conditional density $P(y|x)$. In other words, to each x there corresponds a probabilistic law $P(y|x)$ according to which the selection of y is realized in a random trial.

The existence of such dependences reflects the presence of a stochastic relationship between the vector x and the scalar y. A complete knowledge of these stochastic relations requires the estimation of the conditional density $P(y|x)$. This problem is extremely difficult. However, often in practice (for example, in problems of measurement data processing) it is not necessary to know the function $P(y|x)$ but only one of its characteristics—the conditional mathematical expectation function, i.e., the function which assigns to each x a number $y(x)$ equal to the expectation of the scalar y:

$$y(x) = \int yP(y|x)\, dy.$$

The function $y(x)$ is called the *regression*, and the problem of estimating the conditional mathematical expectation function is referred to as the *problem of regression estimation.*

† In the formulation of the problem one can take into account the differences in the values of errors of the first and second kind. However, this does not change the essence of the problem: the point is that the loss function takes on only a finite number (three) of values.

We shall now consider the statement of this problem. In a certain environment which is characterized by the probability density $P(x)$, a situation x arises randomly and independently. In this environment a transformer acts which assigns to each vector x a number y obtained as a result of the realization of a random trial according to the distribution $P(y|x)$. Neither the properties of $P(x)$ nor those of $P(y|x)$ are known. However, it is known that the regression

$$\bar{y} = y(x)$$

exists.

Based on a random sample of pairs

$$x_1, y_1; \ldots, x_l, y_l,$$

it is required to estimate the regression; in other words, given the class of functions $F(x, \alpha)$, one needs to find a function $F(x, \alpha^*)$ which is closest to the regression $y(x)$.

The problem of estimating the regression is one of the basic problems of applied statistics. The problem of *interpreting the results of direct experiments* can be reduced to the regression problem. Let a lawlike relationship connect the quantity $\bar{y}$ with the vector x by means of a functional relationship

$$\bar{y} = y(x).$$

Let our purpose be to determine the functional relationship $\bar{y} = y(x)$ in the situation when at each point x^* one can conduct a direct experiment to determine this relationship, i.e., direct measurements on the quantity $\bar{y}^* = y(x^*)$ are carried out. However, since the experiment is imperfect, the results of the measurements will determine the true value subject to a certain random error. In other words, at each point x a value $y = y_x$ rather than the value $y(x)$ is obtained. (Here $y_x - y(x) = \xi$ is the experimental error; $M\xi^2 < \infty$.)

It is assumed (and this hypothesis determines the possibility of interpreting experiments) that at no point x is there a systematic error, i.e., the mathematical expectation of the measured function y_x at each fixed point x is equal to the value of the function $y(x)$ at this point:

$$My_x = y(x). \tag{1.7}$$

Moreover, we shall assume that the random variables y_{x_i} and y_{x_j} $(i \neq j)$ are independent.

Under these conditions it is required, on the basis of a finite number of direct experiments, to estimate the function $\bar{y} = y(x)$. Thus the relationship under consideration is the regression (1.7), and the essence of the problem is to estimate regressions based on a sequence of pairs

$$x_1, y_1; \ldots; x_l, y_l.$$

The problem of estimating regression includes the problem of interpreting results of direct experiments. In such problems it is customary to distinguish between two types of experiments: *closed* and *open*. A closed experiment is one in which the probabilistic law $P(x)$—according to which the selection of experimental points is determined—is unknown to the investigator. An open experiment is one in which the law $P(x)$ is known to (and often determined by) the investigator.

The problem of regression estimation reduces to the problem of estimating dependences. Indeed, consider the functional

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy, \tag{1.8}$$

where $P(x, y) = P(y|x)P(x)$. We show that if the regression $\bar{y} = y(x)$ belongs to the class $F(x, \alpha)$ ($y(x) \equiv F(x, \alpha_0)$), then it minimizes the functional (1.8); if, however, the regression does not belong to $F(x, \alpha)$, then the minimum is obtained at the function $F(x, \alpha^*)$ which is closest to the regression. The proximity between the functions $f_1(x)$ and $f_2(x)$ is taken in the L_P^2 metric:

$$\rho_L(f_1(x), f_2(x)) = \left(\int (f_1(x) - f_2(x))^2 P(x)\, dx \right)^{1/2}$$

To show this, denote

$$\Delta F(x, \alpha) = F(x, \alpha) - y(x). \tag{1.9}$$

Then the functional (1.8) can be written in the form

$$I(\alpha) = \int (y - y(x))^2 P(x, y)\, dx\, dy + \int (\Delta F(x, \alpha))^2 P(x)\, dx$$

$$- 2 \int \Delta F(x, \alpha)(y - y(x)) P(x, y)\, dx\, dy.$$

In this expression the third summand is zero, since in view of (1.7),

$$\int \Delta F(x, \alpha)(y - y(x)) P(x, y)\, dx\, dy$$

$$= \int \Delta F(x, \alpha) P(x) \left[\int (y - y(x)) P(y|x)\, dy \right] dx = 0.$$

Thus we have verified that

$$I(\alpha) = \int (y - y(x))^2 P(x, y)\, dx\, dy + \int (F(x, \alpha) - y(x))^2 P(x)\, dx.$$

Since the first summand does not depend on α, the minimum point of $I(\alpha)$ coincides with the minimum point of the second summand, and hence the minimum $I(\alpha)$ is attained on the regression if $y(x) \in F(x, \alpha)$, or at the closest function to it if $y(x) \notin F(x, \alpha)$.

Thus the problem of estimating regression also reduces to the scheme of minimization of expected risk. The special feature of this problem is that the class of functions $Q(z, \alpha)$ admits the following restrictions:

(1) The vector z consists of $n + 1$ coordinates: the coordinate y and n coordinates $x^1, \ldots, x^n$ forming the vector x. However, unlike the case of the pattern recognition problem, the coordinate y as well as function $F(x, \alpha)$ may take on any values in the interval $(-\infty, \infty)$.

(2) The class of functions $Q(z, \alpha)$ is of the form

$$Q(z, \alpha) = (y - F(x, \alpha))^2.$$

The functions $Q(z, \alpha)$ take on arbitrary values on the interval $(0, \infty)$.

§4 The Problem of Interpreting Results of Indirect Experiments

In the preceding section the problem of regression estimation was considered. It was shown that the problem of interpreting the results of direct experiments is reduced to the regression problem. (Recall that in direct experiments the dependence of interest may be measured at any fixed point.) However, it is often the case that the required function $f(t)$ can be measured at no point of t. At the same time some other function $F(x)$ which is connected with $f(t)$ by the operator equation

$$Af(t) = F(x) \tag{1.10}$$

may admit measurements. It is then required, on the basis of the measurements $y_1, \ldots, y_l$ of function $F(x)$ at points $x_1, \ldots, x_l$, to obtain in the class $f(t, \alpha)$ a solution for Equation (1.10). This problem will be called the problem of *interpreting the results of indirect experiments.*

The formation of the problem is as follows: given a continuous operator A which maps in a one-to-one manner the elements $f(t, \alpha)$ of a metric space E_1 into the elements $F(x, \alpha)$ of a metric space E_2, it is required to obtain a solution of the operator equation (1.10) in the class of functions $f(t, \alpha)$ provided the function $F(x)$ is unknown, but the measurements $y_1, \ldots, y_l$ of $F(x)$ at points $x_1, \ldots, x_l$ are given.

As with the interpretation of direct measurements, the measuring experiment of $F(x)$ does not involve systematic error, i.e., $My_{x_i} = F(x_i)$, and the random variables y_{x_i} and y_{x_j} $(i \neq j)$ are independent. Moreover, we shall assume for simplicity that the function $F(x)$ is defined on the interval $[a, b]$. The experiment is open: points x at which measurements of the function $F(x)$ are carried out are randomly and independently distributed on $[a, b]$ according to the uniform distribution.†

† The points x can be defined by any nonvanishing density on $[a, b]$.

The problem of interpreting results of indirect experiments also reduces to the problem of minimizing the expected risk based on empirical data. Indeed, consider the functional

$$I(\alpha) = \int (y - Af(t, \alpha))^2 P(y|x)\, dy\, dx \equiv \int (y - F(x, \alpha))^2 P(y|x)\, dy\, dx.$$

Analogously to the manipulations carried out in Section 3, we obtain

$$\begin{aligned} I(\alpha) &= \int (y - F(x, \alpha))^2 P(y|x)\, dy\, dx \\ &= \int (y - F(x))^2 P(y|x)\, dy\, dx + \int (\Delta F(x, \alpha))^2\, dx \\ &\quad - 2 \int \Delta F(x, \alpha) \left[\int (y - F(x)) P(y|x)\, dy \right] dx, \end{aligned}$$

where

$$\Delta F(x, \alpha) = F(x, \alpha) - F(x).$$

Here the third summand vanishes (as was the case in the preceding section), which implies that the minimum of the functional

$$I(\alpha) = \int (y - Af(t, \alpha))^2 P(y|x)\, dy\, dx \tag{1.11}$$

is attained at the solution $f(t)$ of the operator equation (1.10).

We have thus again arrived at the setup for minimizing the expected risk (1.4) on the basis of empirical data. In this problem the loss function $Q(z, \alpha)$ is such that

(1) the vector z consists of two coordinates y and x, admitting values in the intervals $(-\infty, \infty)$ and $[a, b]$,
(2) the loss function is given by

$$Q(z, \alpha) = (y - Af(t, \alpha))^2.$$

The specific feature of interpreting results of indirect experiments is that we seek a function $f(t, \alpha^*)$ which minimizes the functional (1.11) even though the problem of solving the operator equation

$$Af(t) = F(x), \qquad f(t) \in f(t, \alpha)$$

may be ill posed.

§5 Ill-posed Problems

We say that a solution of the operator equation

$$Af(t) = F(x)$$

is *stable* if the small variation in the right-hand side $F(x) \in F(x, \alpha)$ results in a small change in the solution, i.e., if for any ε a $\delta(\varepsilon)$ can be found such that the inequality

$$\rho_{E_1}(f(t, \alpha_1), f(t)) \leq \varepsilon$$

is valid as long as the inequality

$$\rho_{E_2}(F(x, \alpha_1), F(x)) \leq \delta(\varepsilon)$$

holds. Here the indices E_1 and E_2 denote that the distance is defined in the metrics of spaces E_1 and E_2 respectively (the operator equation (1.10) maps space E_1 into space E_2).

We say that a problem of solving an operator equation is *well posed in the Hadamard sense* if the solution of the equation

(1) *exists*,
(2) is *unique*, and
(3) is *stable*.

A problem of solving an operator equation is considered *ill posed* if the solution of this equation violates at least one of the abovementioned requirements.

Below, in the main portion of the book, we shall confine ourselves to solutions of ill-posed problems of interpreting the results of indirect experiments defined by the Fredholm integral equation of type I:

$$\int_a^b K(t, x) f(t)\, dt = F(x).$$

However, all the results obtained will be valid also for equations defined by any other linear continuous operators.

The necessary background on the theory of solutions of ill-posed problems is given in the Appendix to this chapter.

Thus we shall consider Fredholm's integral equation of type I:

$$\int_0^1 K(x, t) f(t)\, dt = F(x), \tag{1.12}$$

defined by a kernel $K(x, t)$ which is continuous almost everywhere on $0 \leq t \leq 1$, $0 \leq x \leq 1$, and which maps the set of functions $f(t)$ continuous on $[0, 1]$ into the set of functions $F(x)$ continuous on $[0, 1]$.

We shall now show that the problem of solving Equation (1.12) is an ill-posed one. For this purpose we note that a continuous function $G_\nu(x)$ formed by means of the kernel $K(x, t)$,

$$G_\nu(x) = \int_0^1 K(x, t) \sin \nu t \, dt,$$

possesses the property

$$\sup_x G_\nu(x) \underset{\nu \to \infty}{\longrightarrow} 0.$$

Consider the integral equation

$$\int_0^1 K(x, t) \hat{f}(t) \, dt = F(x) + G_\nu(x). \tag{1.13}$$

Since the Fredholm equation is linear, a solution of Equation (1.13) is of the form

$$\hat{f}(t) = f(t) + \sin \nu t,$$

where $f(t)$ is a solution of Equation (1.12). For ν sufficiently large the right-hand sides of Equations (1.12) and (1.13) differ only slightly (by the amount $G_\nu(x)$), while their solutions differ by the amount $\sin \nu t$.

The Fredholm integral equation of type I is one of the basic equations for the problem of interpreting results of indirect experiments. Here are examples of problems connected with a solution of this equation:

EXAMPLE 1 (The Inverse Problem of Spectroscopy). Let the spectrum $F(x)$ be observed using a "real-world" spectroscope. This instrument possesses a finite resolving ability, and the observed spectrum differs in general from the one that would have been observed by means of an ideal spectroscope (i.e., one with an infinitely high resolving power). It is required to calibrate the spectrum obtained by means of the "real-world" spectroscope to the "true" spectrum.

This problem can often be solved. It is known, for example, that the "smoothing" characteristic of certain real-world spectroscopes is of the form

$$K(x,t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-t)^2}{2\sigma^2}\right\}.$$

The observed spectrum $F(x)$ is connected with the true spectrum $f(t)$ by the relation

$$\frac{1}{\sqrt{2\pi}\sigma} \int_0^\infty \exp\left\{-\frac{(x-t)^2}{2\sigma^2}\right\} f(t) \, dt = F(x).$$

The better the instrument (i.e., the smaller the σ), the less the spectral picture is distorted. As $\sigma \to 0$ the characteristic of the apparatus approaches the ideal one:

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-t)^2}{2\sigma^2}\right\} \to \delta(t-x),$$

and hence

$$F(x) \to f(x).$$

However, no matter how poor the real-world spectroscope is, one can in principle derive the actual spectrum from the observed one. For this purpose it is necessary to solve the inverse problem of spectroscopy, i.e., to solve the integral equation

$$\frac{1}{\sqrt{2\pi}\sigma} \int_0^\infty \exp\left\{-\frac{(x-t)^2}{2\sigma^2}\right\} f(t)\,dt = F(x),$$

utilizing the empirical data $y_1, \ldots, y_l$ in place of the function $F(x)$.

EXAMPLE 2 (The Problem of Identifying Linear Objects). It is known that dynamic properties of linear homogeneous objects with one output are completely described by the pulse-transfer (weight) function $f(\tau)$. The function $f(\tau)$ is the reaction of the object to a unit pulse served at the system at time $\tau = 0$.

Knowing this function one can compute the reaction of the object to any disturbance $x(t)$ using the formula

$$y(t) = \int_0^t x(t-\tau) f(\tau)\,d\tau.$$

Thus the determination of the dynamic characteristics of an object reduces to the determination of the weight function $f(\tau)$.

It is known that for a linear homogeneous object the Wiener–Hopf equation

$$\int_0^\infty R_{xx}(t-\tau) f(\tau)\,d\tau = R_{yx}(t) \tag{1.14}$$

is valid. Equation (1.14) connects the autocorrelation function $R_{xx}(t)$ of a stationary random process at the input of the object with the weight function $f(\tau)$ and joint correlation function of the input and output signals $R_{yx}(t)$. Thus the problem of identifying a linear object involves the determination of a weight function based on the known autocorrelation function of the input signal and the measured (observed) joint correlation function of the input and output signals, i.e., it is the problem of solving the integral equation (1.14) on the basis of empirical data.

EXAMPLE 3 (The Problem of Estimating Derivatives). Let the measurements of smooth function $F(x)$ at l points of the interval $[0, 1]$ be given. The points at which the measurements were taken are distributed randomly and independently according to a uniform distribution. It is required to estimate on $[0, 1]$ the derivative $f(x)$ of the function $F(x)$.

It is easy to see that the problem is reduced to solving the Volterra integral equation of type I,

$$\int_0^x f(t)\,dt = F(x) - F(0),$$

under the condition that the l measurements $y_1, \ldots, y_l$ of the function $F(x)$ carried out at points $x_1, \ldots, x_l$ are known. Equivalently it reduces to the solution of the type-I Fredholm equation (under the same conditions),

$$\int_0^1 \theta(x - t)f(t)\,dt = F(x) - F(0),$$

where

$$\theta(z) = \begin{cases} 1 & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

In a more general case when the kth derivative is to be estimated, the following integral equation must be solved:

$$\int_0^1 \frac{(x - t)^{k-1}}{(k-1)!}\theta(x - t)f(t)\,dt = F(x) - \sum_{j=0}^{k-1} \frac{F^{(j)}(0)}{j!},$$

where in place of function $F(x)$ the empirical data $y_1, \ldots, y_l$ are used. Here $F^{(j)}(0)$ is the value of the jth derivative at zero.

§6 Accuracy and Confidence of Risk Minimization Based on Empirical Data

We have thus considered three basic problems of estimating dependences from the empirical data: pattern recognition, regression estimation and interpretation of indirect experiments. They are all based on the same general setup: the model of minimizing the expected risk based on empirical data. In other words, it is required to find α^* which minimizes the functional

$$I(\alpha) = \int Q(z, \alpha)P(z)\,dz,$$

where the density $P(z)$ is unknown but a random independent sample $z_1, \ldots, z_l$ of size l is given.

Moreover, for all these problems the same structure of the loss function,

$$Q(z, \alpha) = (y - F(x, \alpha))^2$$

was chosen. Thus in all cases it is required to obtain a function $F(x, \alpha^*)$ which minimizes the functional

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy, \tag{1.15}$$

where the density $P(x, y)$ is unknown but a sample $x_1, y_1; \ldots; x_l, y_l$ obtained from random and independent trials according to this density is given. Actually we have distinguished between two variant formulations of the problem of regression estimation: the case when the density $P(x)$ is unknown (a closed experiment) and the case when $P(x)$ is known (an open experiment). But these two formulations do not differ fundamentally, the main point being that the joint density $P(x, y)$ is unknown in both cases.

We have established that various problems of estimating dependences differ as the loss functions for risk minimization differ, and that in each problem it is the parameter α which yields the exact minimum for the corresponding functional determines the required functional relationship. However, to obtain the exact minimum of the functional (1.15) from a sample of a fixed size is generally an insoluble problem, since a sample is only a "realization" of the underlying distribution law and is in no way equivalent to it. Therefore one should consider the problem of determining, from a sample of a fixed size, a function which yields the value of the functional "close" to the minimal one rather than the exact minimum of the functional (1.15).

Moreover, one cannot guarantee that a value "close" to the minimum will be obtained unconditionally, but only with a certain probability (since, given any density, there is a certain probability that the sample obtained in random trials will consist of l pairs of elements x, y repeated l times). Thus the preassigned *accuracy* of minimizing the expected risk (1.15) can be obtained from a sample of a fixed size only with a certain *confidence*.

We say that the value of the functional $I(\alpha^*)$ is $\varkappa$-close to the minimal $(\min_\alpha I(\alpha))$ if the inequality

$$I(\alpha^*) - \min_\alpha I(\alpha) \leq \varkappa$$

is fulfilled. Now let an algorithm A which determines the value of parameter α^* from a sample of size l be given. Since the sample is random, this algorithm determines a random value of the parameter α^* to which the random number $I(\alpha^*)$ corresponds. We say that an algorithm A yields with α confidence level $1 - \eta$ a value of the functional $I(\alpha)$ which is $\varkappa$-close to the minimal if for any given $0 < \eta < 1$ the inequality

$$P\left\{I(\alpha^*) - \min_\alpha I(\alpha) > \varkappa\right\} < \eta$$

is valid. When solving problems of expected-risk minimization our purpose is to obtain algorithms which for a sample of a fixed size and with a given confidence level will determine functions yielding the value of functional $I(\alpha)$ that is, closest to the minimum.

§7 The Accuracy of Estimating Dependences on the Basis of Empirical Data

At the end of the preceding section the purpose of our investigation was formulated: to find algorithms which guarantee that the risk closest to the minimal will be attained. This book is devoted to the construction and justification of such algorithms. However, when formulating the goal of the investigation the problem was in essence replaced by another. Indeed the initial goal was to estimate functional dependences. In Sections 2, 3, and 4 it was shown that a function which yields the exact minimum of a corresponding functional of the expected risk determines the required dependence. On the other hand, to obtain an exact minimum from a sample of a fixed size is an unrealistic problem. It was therefore suggested to search for a function which yields a value of the expected risk close to the minimal.

However, it does not follow at all that close functions will correspond to close values of the functionals. Determining the value of a functional which is close to the minimal one is in general a different problem. Therefore, before solving the problem of estimating functional dependences from empirical data using the method of minimizing expected risk, it is necessary to find out whether this substitution of the problem will be adequate, i.e., whether the closeness of the functionals assures the closeness of the functions.

In order to begin an investigation in this direction, it is first necessary to define precisely the "closeness" of functions. Unlike the closeness of functionals, which can be defined naturally as the distance between two points on the real line (which represent the values of these functionals), the closeness between functions has to be defined as the distance between two elements of a function space.

There are various methods of metrization (introduction of the notion of distance) in functional analysis. We shall utilize two such metrics: a weighted mean-square deviation and a uniform deviation. The distance between two functions $f_1(x)$ and $f_2(x)$ in the mean-square sense with weight $P(x)$ (the L_P^2 metric) is defined by the functional

$$\rho_L(f_1(x), f_2(x)) = \left(\int (f_1(x) - f_2(x))^2 P(x)\, dx \right)^{1/2},$$

where $P(x)$ is a nonnegative function such that $\int P(x)\,dx = 1$. The distance in the uniform deviation sense (the C metric) is defined by the functional

$$\rho_C(f_1(x), f_2(x)) = \sup_x |f_1(x) - f_2(x)|.$$

Thus two functions are close in the L_P^2 metric if

$$\left(\int (f_1(x) - f_2(x))^2 P(x)\,dx\right)^{1/2} \leq \varkappa, \tag{1.16}$$

and are close in the C metric if

$$\sup_x |f_1(x) - f_2(x)| \leq \varkappa. \tag{1.17}$$

Note that the requirement of uniform closeness (1.17) is stronger than that of mean-square closeness. The inequality (1.17) implies (1.16), but the converse is generally not true.

Thus we shall use the notion of closeness (proximity) in the following senses:

(1) Closeness of qualities of functions (values of functionals).
(2) Closeness of functions in the L_P^2 metric.
(3) Closeness of functions in the C metric.

The choice of the closeness measure is determined by the nature of the problem and not formally.

How is closeness defined in various problems of estimating dependences?

In a pattern recognition problem it is required, in a given class of indicator functions, to find a function which minimizes the probability of erroneous classification (i.e., it is required to minimize a functional). Therefore it is natural here to consider two functions to be close if their "qualities" are close; here the proximity is defined by the proximity of the functionals.

In the case of regression estimation, the problem is to find a function which is close to the regression rather than to minimize a functional. In this problem the proximity is defined by means of L_P^2 or C metrics, depending on how the estimated function is to be used later on.

For example, consider the problem of estimating the regression $\bar{y} = y(x)$ in the setup for interpreting direct experiments. The estimated dependence $\bar{y} = F(x, \alpha^*)$ is to be used to forecast the value of $\bar{y}$ for different values of the situation x. The accuracy of the forecast for a given x is natural to measure by the quantity

$$(y(x) - F(x, \alpha^*))^2.$$

The overall accuracy of the forecast based on the estimated function is often measured as the average accuracy with respect to the measure of the set x,

i.e., by the quantity

$$\rho_L(y(x), F(x, \alpha)) = \left(\int (y(x) - F(x, \alpha))^2 P(x)\, dx\right)^{1/2}.$$

In other words, the proximity is determined here by the L_P^2 metric.

There are, however, problems where the proximity in the L_P^2 metric is not sufficient. Let, for example, a quantity $\bar{y}$ be functionally related to technological parameters x. It is required to obtain a vector of parameters x^* which will yield the maximum of $\bar{y}$. This problem is solved according to the following scheme: first the functional relationship $\bar{y} = y(x)$ is estimated, and then a value x^* is sought which yields the maximum of the estimated function. However, if in this case a function $F(x, \alpha^*)$ close to the actual one in the L_P^2 metric is chosen, then the situation shown in Figure 1 may occur.

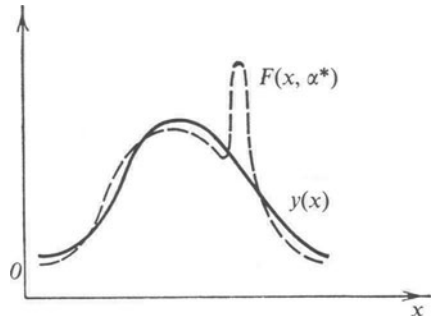

Figure 1

The estimated function may approximate the true function sufficiently well almost everywhere except for a set x of small measure where a large outlier occurs. The maximum of the estimated function, however, does not reflect the point yielding the maximum of $\bar{y}$, but the outlier of the estimated function.

In order to exclude such a situation it is necessary that the estimated function should approximate the true one uniformly over the whole domain of the definition of the function, i.e., in the metric C

$$\rho_C(y(x), F(x, \alpha)) = \sup_x |y(x) - F(x, \alpha)|.$$

Thus in problems of regression estimation, closeness both in the L_P^2 metric and in the C metric are used.

In the problem of interpreting data from indirect experiments, two notions of closeness are also used: closeness in the L_P^2 metric (which is L_P^2 with the weight $P(x) \equiv 1$)

$$\rho_L(f(t, \alpha_1), f(t, \alpha_2)) = \left(\int (f(t, \alpha_1) - f(t, \alpha_2))^2\, dt\right)^{1/2},$$

and in the C metric

$$\rho_C(f(t, \alpha_1), f(t, \alpha_2)) = \sup_t |f(t, \alpha_1) - f(t, \alpha_2)|.$$

As in the case of a regression problem, the choice of the metric is determined by the manner in which the estimated function is further utilized.

§8 Special Features of Problems of Estimating Dependences

We have thus established that all three problems of estimating dependences are reduced to the same setup—the problem of minimizing the expected risks—and that only an approximate solution of the latter problem is possible on the basis of empirical data. The question arises: Does an approximate solution assure the required closeness of the dependence obtained to the actual one?

The answer to this question depends on the problem at hand. For a pattern recognition problem the answer is unequivocally yes by definition (since according to the statement of the problem it is required to find a function which yields a value of the functional close to the minimal one).

In the case of regression estimation the answer is not as clearcut. It can be easily shown that if we interpret the proximity of functions in the L_P^2 sense, then the proximity of a functional to a minimal one yields the proximity of the function obtained to the regression. A proof of this assertion follows directly from the identity

$$\int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy$$

$$= \int (y - y(x))^2 P(x, y)\, dx\, dy + \int (y(x) - F(x, \alpha))^2 P(x)\, dx,$$

where $\bar{y} = y(x)$ is the regression and $F(x, \alpha)$ is an arbitrary function belonging to a given class. However, the proximity of a functional to the minimal one does not in any way imply the proximity in the C sense of the corresponding function to the regression. To assure such a proximity it is not sufficient simply to minimize the functional. It is necessary that certain special requirements be satisfied.

Finally for the problem of interpreting the results of indirect experiments the proximity of a functional to the minimal one does not assure the proximity of the estimated function to the actual one, either in the L_P^2 or in the C metrics. The basic difficulty in solving this problem is that the solution of the corresponding operator equation may be an ill-posed problem, and in this case functions which yield values of the functional close to the minimal

one may differ significantly from the desired solution. Therefore the main problem here is to determine what additional conditions should be imposed on the chosen solution in order that the proximity of the functional obtained to the minimal one will result in the proximity of the solution to the desired function.

Thus, in spite of the fact that in all the problems of estimating dependences the functions yielding an exact minimum of the functional determine a solution, an approximate minimization does not always result in achieving this goal. Therefore, before applying a specific method of minimizing expected risks based on empirical data, it is necessary to make sure that the minimization method assures an approximation to the desired solution.

In subsequent chapters various methods of minimization of expected risks based on empirical data are considered. They are all studied in connection with each of the specific problems of estimating dependences.

Appendix to Chapter 1

Methods for Solving Ill-posed Problems

§A1 The Problem of Solving an Operator Equation

We say that two sets of elements, $\mathcal{M}$ and $\mathcal{N}$, are *functionally dependent* if given any element of $f \in \mathcal{M}$ there corresponds a unique element $F \in \mathcal{N}$.

This functional dependence is called a function if the sets $\mathcal{M}$ and $\mathcal{N}$ are sets of numbers; it is called a functional if $\mathcal{M}$ is a set of functions and $\mathcal{N}$ is a set of numbers, and it is called an operator if both sets are sets of functions.

Each operator A uniquely maps elements of the set $\mathcal{M}$ into elements of the set $\mathcal{N}$. This is denoted by the equality

$$A\mathcal{M} = \mathcal{N}.$$

In a collection of operators we shall single out those which realize a one-to-one mapping of $\mathcal{M}$ into $\mathcal{N}$. For these operators the problem of solving the operator equation

$$Af(t) = F(x) \tag{A.1}$$

can be considered as the problem of finding an element $f(t)$ in $\mathcal{M}$ to which an element $F(x)$ corresponds in $\mathcal{N}$.

For operators which realize a one-to-one mapping of elements $\mathcal{M}$ into $\mathcal{N}$ and a function $F(x) \in \mathcal{N}$ there exists a unique solution of the operator equation (A.1). However, to obtain a method for solving an operator equation of such generality is a hopeless task. Therefore we shall investigate operator equations with continuous operators only.

Let the elements $f \in \mathcal{M}$ belong to a metric space E_1 with metric $\rho_1(\cdot)$, and the elements $F \in \mathcal{N}$ belong to a metric space E_2 with metric $\rho_2(\cdot)$. An operator A is called *continuous* if "close" elements (with respect to metric ρ_1) in E_1 are mapped into "close" elements (with respect to metric ρ_2) in E_2.

We shall consider an operator equation defined by a continuous operator which maps in a one-to-one manner $\mathcal{M}$ into $\mathcal{N}$. The solution of such an operator equation exists and is unique, i.e., there exists the inverse operator A^{-1} from $\mathcal{N}$ into $\mathcal{M}$:

$$\mathcal{M} = A^{-1}\mathcal{N}.$$

The basic problem is whether the inverse operator is continuous.

If operator A^{-1} is continuous, then close preimages will correspond to close functions in $\mathcal{N}$, i.e., the solution of the operator equation will be stable. If, however, the inverse operator is not continuous, then the solution of the operator equation will generally be unstable. In the latter case, in view of Hadamard's definition (Chapter 1, Section 5) the problem of solving an operator equation is considered ill posed. It turns out that in many important cases, for example, for completely continuous operators A, the inverse operator A^{-1} is not continuous and hence the problem of solving the corresponding operator equation is ill posed.

Definition. We say that a linear operator A defined in a linear normed space E_1 with the range of values in a linear normed space E_2 is *completely continuous* if it maps any bounded set in space E_1 into a compact set of the space E_2, i.e., if each bounded infinite sequence in E_1

$$f_1, f_2, \ldots, f_i, \ldots, \qquad \|f_j\| \leq c, \tag{A.2}$$

(here $\|f_j\|$ is the norm in E_1) is mapped in E_2 into a sequence

$$Af_1, \ldots, Af_i, \ldots, \tag{A.3}$$

such that a convergent subsequence

$$Af_{i_1}, \ldots, Af_{i_k}, \ldots \tag{A.4}$$

can be extracted from it.

We shall show that if the space E_1 contains bounded noncompact sets, then the inverse operator A^{-1} for a continuous operator A need not be continuous. Indeed, consider in E_1 a bounded noncompact set. Select in this set an infinite sequence (A.2) such that no subsequence of it is convergent. An infinite sequence (A.3) from which a convergent subsequence (A.4) may be selected (since operator A is absolutely continuous) corresponds in E_2 to this sequence. If the operator A^{-1} were continuous, then a convergent sequence

$$f_{i_1}, \ldots, f_{i_k}, \ldots, \tag{A.5}$$

would correspond to the sequence (A.4) in E_1 which will be a subsequence of (A.2). This, however, contradicts the choice of (A.2).

Thus the problem of solving an operator equation defined by a completely continuous operator is an ill-posed problem. In the main part of this book we shall consider linear integral operators

$$Af = \int_a^b K(t, x) f(t)\, dt \tag{A.6}$$

with a continuous kernel $K(t, x)$ in the domain $a \leq t \leq b$, $a \leq x \leq b$. Operators (A.6) are completely continuous from $C[a, b]$ into $C[a, b]$. The proof of this fact can be found in all texts on functional analysis (see, for example, [28]).

§A2 Problems Well Posed in Tihonov's Sense

The problem of solving the operator equation

$$Af = F$$

is called *well posed (correct) in Tihonov's sense* on the set $\mathcal{M}' \subset \mathcal{M}$, and the set $\mathcal{M}'$ is called the *set (class) of correctness*, provided:

(1) the exact solution of the problem exists for each $F \in \mathcal{N} = A\mathcal{M}'$ and belongs to $\mathcal{M}'$;
(2) the solution belonging to $\mathcal{M}'$ is unique for any $F \in A\mathcal{M}' = \mathcal{N}'$;
(3) solutions belonging to $\mathcal{M}'$ are stable with respect to $F \in \mathcal{N}'$.

If $\mathcal{M}' = \mathcal{M}$ and $\mathcal{N}' = \mathcal{N}$ then correctness in Tihonov's sense corresponds to correctness in Hadamard's sense. The meaning of Tihonov's correctness is that correctness can be achieved by restricting the set of solutions $\mathcal{M}$ to a class of correctness $\mathcal{M}'$.

The following lemma shows that if we narrow the set of solutions $\mathcal{M}$ to a compact set $\mathcal{M}'$, then it constitutes a correctness class.

Lemma. *If a continuous one-to-one operator A is defined on a compact $\mathcal{M}' \subset \mathcal{M}$, then the inverse operator A^{-1} is continuous on the set $\mathcal{N}' = A\mathcal{M}'$.*

PROOF. Choose an arbitrary element $F_0 \in \mathcal{N}'$ and an arbitrary sequence convergent to it:

$$\{F_n\} \subset \mathcal{N}', \qquad F_n \underset{n \to \infty}{\longrightarrow} F_0.$$

It is required to verify the convergence

$$f_n = A^{-1}F_n \underset{n \to \infty}{\longrightarrow} A^{-1}F_0 = f_0.$$

Since $\{f_n\} \subset \mathcal{M}'$, and $\mathcal{M}'$ is a compact set, the limit points of the sequence $\{f_n\}$ belong to $\mathcal{M}'$. Let f_0 be such a limit point. Since f_0 is a limit point, there exists a sequence $\{f_{n_k}\}$ convergent ot it, to which there corresponds a

sequence $\{F_{n_k}\}$ convergent to F_0. Therefore, approaching the limit in the equality

$$Af_{n_k} = F_{n_k}$$

and utilizing the continuity of the operator A, we obtain

$$Af_0 = F_0.$$

Since the operator A^{-1} is unique, we have $A^{-1}F_0 = f_0$, which implies the uniqueness of the limit point of the sequence $\{f_{n_k}\}$. It remains to verify that the whole sequence $\{f_{n_k}\}$ converges to f_0. Indeed, if the whole sequence is not convergent to f_0, one could find a neighborhood of the point f_0 outside of which there would be infinitely many members of the sequence $\{f_{n_k}\}$. Since $\mathcal{M}'$ is compact, this sequence possesses a limit point f'_0 which, by what was proven above, coincides with f_0. This, however, contradicts the assumption that the selected sequence lies outside a neighborhood of point f_0. The lemma is thus proved. □

Hence correctness in Tihonov's sense on a compactum $\mathcal{M}'$ follows from the conditions of the existence and uniqueness of a solution of an operator equation. The third condition (the stability of the solution) is automatically satisfied. This fact is essentially the basis for all constructive ideas for solving ill-posed operator equations. We shall consider one of them.

§A3 The Regularization Method

The regularization method was proposed by A. N. Tihonov in 1963.

It is required to solve the operator equation

$$Af = F, \tag{A.7}$$

defined by a continuous one-to-one operator A acting from $\mathcal{M}$ into $\mathcal{N}$. Let the solution of (A.7) exist.

We introduce a lower semicontinuous functional $\Omega(f)$, which we shall call the *stabilizer* and which possesses the following three properties:

(1) the solution of the operator equation belongs to the domain of definition $D(\Omega)$ of functional $\Omega(f)$;
(2) on the domain of the definition functional $\Omega(f)$ admits real-valued nonnegative values;
(3) the sets

$$\mathcal{M}_c = \{f : \Omega(f) \leq c\}, \qquad c \geq 0,$$

are all compact.

The idea of regularization is to find a solution for (A.7) as an element minimizing a certain functional. It is not the functional

$$\rho = \rho_2(Af, F)$$

(this problem would be equivalent to the solution of Equation (A.7) and therefore would also be ill posed), but an "improved" functional

$$R_\gamma(\hat{f}, F) = \rho_2^2(A\hat{f}, F) + \gamma\Omega(\hat{f}), \qquad \hat{f} \in D(\Omega), \tag{A.8}$$

with *regularization parameter* $\gamma > 0$. The problem of minimizing the functional (A.8) is stable, i.e., to the close functions F and F_δ (where $\rho_2(F, F_\delta) \leq \delta$) there correspond close elements f^γ and f_δ^γ which minimize the functionals $R_\gamma(f, F)$ and $R_\gamma(f, F_\delta)$.

The problem is to determine a relationship between δ and γ such that the sequence of solutions f_δ^γ of regularized problems $R_\gamma(f; F_\delta)$ will converge as $\delta \to 0$ to the solution of the operator equation (A.7). The following theorem establishes these relations.

Theorem A.1. *Let E_1 and E_2 be metric spaces, and let there exist for $F \in \mathcal{N}$ a solution of Equation* (A.7) *for $f \in D(\Omega)$. Then if in place of an exact right-hand side F of Equation* (A.7), *approximations*† *$F_\delta \in E_2$ are known such that $\rho_2(F, F_\delta) \leq \delta$ and the values of parameter γ are chosen in such a manner that*

$$\gamma(\delta) \to 0 \quad \text{for } \delta \to 0, \qquad \lim_{\delta \to 0} \frac{\delta^2}{\gamma(\delta)} \leq r < \infty, \tag{A.9}$$

it follows that the elements $f_\delta^{\gamma(\delta)}$ minimizing the functionals $R_{\gamma(\delta)}(f, F_\delta)$ on $D(\Omega)$ converge to the exact solution f as $\delta \to 0$.

Proof. The proof of the theorem utilizes the following fact: for any fixed $\gamma > 0$ and an arbitrary $F \in \mathcal{N}$ an element $f^\gamma \in D(\Omega)$ exists which minimizes the functional $R_\gamma(f, F)$ on $D(\Omega)$.

Let γ and δ satisfy the relation (A.9). Consider a sequence of elements $f_\delta^{\gamma(\delta)}$ minimizing $R_{\gamma(\delta)}(f, F_\delta)$, and show that the convergence

$$f_\delta^{\gamma(\delta)} \underset{\delta \to 0}{\longrightarrow} f$$

is valid. By definition we have

$$\begin{aligned} R_{\gamma(\delta)}(f_\delta^{\gamma(\delta)}, F_\delta) &\leq R_{\gamma(\delta)}(f, F_\delta) = \rho_2^2(Af, F_\delta) + \gamma(\delta)\Omega(f) \\ &\leq \delta^2 + \gamma(\delta)\Omega(f) = \gamma(\delta)\left(\Omega(f) + \frac{\delta^2}{\gamma(\delta)}\right). \end{aligned}$$

Taking into account that

$$R_{\gamma(\delta)}(f_\delta^{\gamma(\delta)}, F_\delta) = \rho_2^2(Af_\delta^{\gamma(\delta)}, F_\delta) + \gamma(\delta)\Omega(f_\delta^{\gamma(\delta)}),$$

† The elements F_δ need not belong to the set $\mathcal{N}$.

we conclude

$$\Omega(f_\delta^{\gamma(\delta)}) \leq \Omega(f) + \frac{\delta^2}{\gamma(\delta)},$$

$$\rho_2^2(Af_\delta^{\gamma(\delta)}, F_\delta) \leq \gamma(\delta)\left(\Omega(f) + \frac{\delta^2}{\gamma(\delta)}\right).$$

Since the conditions (A.9) are fulfilled, all the elements of the sequence $f_\delta^{\gamma(\delta)}$ for a $\delta > 0$ sufficiently small belong to a compactum $\mathcal{M}_{c^*}$, where $c^* = \Omega(f) + r + \varepsilon$, $\varepsilon > 0$, and their images $F_\delta^{\gamma(\delta)} = Af_\delta^{\gamma(\delta)}$ are convergent:

$$\begin{aligned}\rho_2(F_\delta^{\gamma(\delta)}, F) &\leq \rho_2(F_\delta^{\gamma(\delta)}, F_\delta) + \delta \\ &\leq \delta + \sqrt{\delta^2 + \gamma(\delta)\Omega(f)} \underset{\delta \to 0}{\longrightarrow} 0.\end{aligned}$$

This implies, in view of the lemma, that their preimages

$$f_\delta^{\gamma(\delta)} \to f \quad \text{for } \delta \to 0$$

are also convergent, q.e.d. □

In a Hilbert space the functional $\Omega(f)$ may be chosen to be equal to $\|f\|^2$ for a linear operator A. Although the sets $\mathcal{M}_c$ are (only) weakly compact in this case, the convergence of regularized solutions—in view of the properties of Hilbert spaces—will be, as shown below, a strong one. Such a choice of a regularizing functional is convenient also because its domain of definition $D(\Omega)$ coincides with the whole space E_1. However, in this case the conditions imposed on the parameter γ are more rigid than in the case of Theorem A.1: γ should converge to zero slower than δ^2.

Thus the following theorem is valid.

Theorem A.2. *Let E_1 be a Hilbert space and $\Omega(\hat{f}) = \|\hat{f}\|^2$. Then for $\gamma(\delta)$ satisfying the relations* (A.9) *with $r = 0$, the regularized elements $f_\delta^{\gamma(\delta)}$ converge as $\delta \to 0$ to the exact solution f in the metric of the space E_1.*

PROOF. It is known from the geometry of Hilbert spaces that the sphere $\|f\|^2 \leq c$ is a weak compactum and that from the properties of weak convergence of elements f_i to the element f and convergence of the norms $\|f_i\|$ to $\|f\|$ there follows the strong convergence

$$\|f_i - f\| \underset{i \to \infty}{\longrightarrow} 0.$$

Moreover it follows from the weak convergence $f_i \to f$ that

$$\|f\| \leq \varliminf_{i \to \infty} \|f_i\|. \tag{A.10}$$

Utilizing these properties of Hilbert spaces, we shall now prove the theorem.

First we note that for a weak convergence in the space E_1 the preceding theorem is valid: $f_\delta^{\gamma(\delta)}$ converges weakly to f as $\delta \to 0$. Therefore in view of (A.10) the inequality

$$\|f\| \leq \underline{\lim}_{\delta \to 0} \|f_\delta^{\gamma(\delta)}\|$$

is valid. On the other hand, taking into account that $\Omega(f) = \|f\|^2$ and that $r = 0$, we obtain

$$\overline{\lim}_{\delta \to 0} \|f_\delta^{\gamma(\delta)}\|^2 \leq \lim_{\delta \to 0} \left(\|f\|^2 + \frac{\delta^2}{\gamma(\delta)} \right) = \|f\|^2.$$

Hence the convergence of the norms is valid:

$$\|f_\delta^{\gamma(\delta)}\| \underset{\delta \to 0}{\longrightarrow} \|f\|,$$

and along with it the validity of weak convergence implies, in view of the properties of Hilbert spaces, the strong convergence

$$\|f_\delta^{\gamma(\delta)} - f\| \underset{\delta \to 0}{\longrightarrow} 0,$$

q.e.d. □

The theorems presented above are basic in regularization theory. Using these theorems the feasibility of solving ill-posed problems is established. However, for solving practical problems the question of convergence of a sequence of regularized solutions is not the most topical. Usually the right-hand side of an operator equation is defined with finite accuracy δ, and the problem is to determine the value of the constant of regularization $\gamma(\delta)$ which will assure the best approximation to the desired solution. In this situation the assertions of Theorems A.1 and A.2, in which the value of γ is determined only up to a constant r (and only for δ sufficiently small), are obviously insufficient.

At present there are no reliable methods for choosing the constant of regularization. However, there are numerous examples where for a suitable choice of constant γ sufficiently good approximations to solutions of ill-posed problems can be obtained.

A detailed treatment of the theory of ill-posed problems is given in the monograph [56].

Chapter 2

Methods of Expected-Risk Minimization

§1 Two Approaches to Expected-Risk Minimization

There are two approaches to solving the problem of minimizing the expected risk

$$I(\alpha) = \int Q(z, \alpha)P(z)\,dz \tag{2.1}$$

on the basis of empirical data

$$z_1, \ldots, z_l. \tag{2.2}$$

The first approach is connected with the idea of constructing, from the sample (2.2) and the function $Q(z, \alpha)$, an *empirical functional*

$$I_{\text{emp}}(\alpha) = \Phi(Q(z_1, \alpha), \ldots, Q(z_l, \alpha); z_1, \ldots, z_l), \tag{2.3}$$

i.e., a functional which does not depend on the unknown probability density $P(z)$. Unlike (2.1), the functional (2.3) can be minimized. We choose its minimum point as that of the initial functional (2.1). This is called the *method of minimizing empirical functionals.*

The basic problem encountered in studying this method is to determine the error size for each type of approximation (2.3) and to obtain an approximation for the functional (2.1) in terms of the empirical functional (2.3) so as to assure the determination of a function which will yield the value of the functional (2.1) close to the minimum.

The second approach connects the determination of the minimum of the functional (2.1) with the use of the iterative procedure

$$\alpha(i) = \alpha(i - 1) + \gamma(i)S(i, z_i). \tag{2.4}$$

According to this procedure the improvement of the vector of parameters α at the ith step is determined by the size $\gamma(i)$ and the direction $S(i, z_i)$ of the ith step. It turns out that if one chooses the direction $S(i, z_i)$ in such a manner that at each step the inequality†

$$(\nabla_\alpha I(\alpha(i-1)))^T MS(i, z) \geq \delta > 0 \tag{2.5}$$

is satisfied, where $\nabla_\alpha I(\alpha)$ is the gradient with respect to α of the functional (2.1), $MS(i, z)$ is the mathematical expectation of the direction of the ith step, then under some additional conditions which restrict the growth of the vector $S(i, z)$ (for example, by means of the function $|z|$) and that of the size of the step $\gamma(i)$ (by assuming that $\sum_{i=1}^{\infty} \gamma^2(i) < \infty$ but at the same time $\sum_{i=1}^{\infty} \gamma(i) = \infty$), the procedure (2.4) and the random sample $z_1, \ldots, z_l, \ldots$ generate a sequence $\alpha(i)$ which converges to the vector of parameters α_0, yielding the minimum of the functional (2.1) (cf. [45]).

The iterative procedure (2.4) is a development of gradient methods of search for minima. Indeed, if the density $P(z)$ were known, then one could under certain conditions compute the gradient

$$\nabla_\alpha I(\alpha) = \int \nabla_\alpha Q(z, \alpha) P(z)\, dz. \tag{2.6}$$

Then the descent procedure would be the following rule:

$$\alpha(i) = \alpha(i-1) - \gamma(i) \nabla_\alpha I(\alpha(i-1)). \tag{2.7}$$

The procedure (2.4) differs from (2.7) in that at each step, it chooses a direction of motion that is "on the average, approximately the same as along the gradient" rather than the direction of the gradient itself. The inequality (2.5) formalizes the expression "on the average, approximately in the same direction".

Thus the basic result of the theory of iterative methods is that, even under quite general conditions on the direction of motion and the size of a step, the iterative procedures (2.4) achieve their purpose. However, due to the very universality of the iterative procedure, the determination of the value of a functional close to the minimal is assured only asymptotically. For solving problems of minimizing the expected risk on the basis of a sample of fixed size, iterative methods are of little use. Therefore we will not consider these methods. Solutions of the problem of minimizing the functional (2.1) on the basis of empirical data (2.2) will therefore be associated with the construction of empirical functional (2.3) and its subsequent minimization.

† Here and below, a vector is assumed to be a column vector and T denotes transposition.

§2 The Problem of Large Deviations

Our purpose is to construct a method which will assure with a given probability the determination of a function yielding the value of functional

$$I(\alpha) = \int Q(z, \alpha)P(z)\,dz$$

which is close to minimal (where the density $P(z)$ is unknown but the sample $z_1, \ldots, z_l$ is given).

Without utilizing prior information this problem cannot be solved. Indeed, consider one of the simplest problems of estimating relationships based on empirical data. It is required to minimize the functional

$$I'(\alpha) = \int (t - \alpha)^2 P(t)\,dt, \tag{2.8}$$

provided $P(t)$ is unknown (it is known only that a variance exists) but a random independent sample $t_1, \ldots, t_l$ is given. The minimum of the functional (2.8) is attained at

$$\alpha = \int tP(t)\,dt. \tag{2.9}$$

Thus the problem is to find for an unknown density $P(t)$ a method which will assure, with a given probability, a sufficiently accurate estimator of the mean based on a sample of a fixed size l.

It turns out that without *a priori* information on the density $P(t)$ one cannot obtain a guaranteed estimator of the mean. Indeed, let the random variable t take on the two values 0 and K, and let $P(t = 0) = 1 - \varepsilon$ and $P(t = K) = \varepsilon$. Assume now that ε is so small that with a high probability $1 - \delta$ the random independent sample $t_1, \ldots, t_l$ consists solely of zeros and hence the value of the empirical mean

$$\alpha_{\text{emp}} = \frac{1}{l} \sum_{i=1}^{l} t_i$$

is zero. (The probability of this event is $(1 - \varepsilon)^l = 1 - \delta$.) On the other hand, the mathematical expectation of the random variable t equals

$$Mt = 0(1 - \varepsilon) + K\varepsilon = K\varepsilon,$$

and depending on the value of K may admit arbitrary values including large ones (for example, when $K = 1/\varepsilon^2$). Thus in our example, in spite of the fact that almost any (random) value of the empirical mean based on a sample of size l is zero, one can come to no reliable conclusions concerning the value of the mathematical expectation. This is because the product $K\varepsilon$ may be large even for small ε. In other words the distribution of the random

variable t is such that a large value K is concentrated on a "small measure" ε. Such situations are termed in statistics "large deviations" of random variables.

When, then, can one reach a reliable conclusion about the value of the mathematical expectation, based on the value of the empirical mean? The answer to this question follows from Chebyshev's inequality. According to this inequality the probability of deviation of a random variable t from its expected value Mt is bounded by

$$P\{|t - Mt| \geq \sigma\varkappa\} \leq \frac{1}{\varkappa^2},$$

where σ^2 is the variance of the variable t. Consider now the random variables

$$\xi = \frac{1}{l}\sum_{i=1}^{l} t_i,$$

where $t_1, \ldots, t_l$ is a random independent sample of size l. Observe that

$$M\xi = Mt, \qquad \sigma_\xi = \frac{\sigma}{\sqrt{l}}.$$

Chebyshev's inequality for this variable becomes

$$P\left\{\left|\frac{1}{l}\sum_{i=1}^{l} t_i - Mt\right| \geq \frac{\sigma}{\sqrt{l}}\right\} \leq \frac{1}{\varkappa^2}. \tag{2.10}$$

We write (2.10) in a different form. Denote the right-hand side by η, i.e., $1/\varkappa^2 = \eta$, or $\varkappa = 1/\sqrt{\eta}$. In this notation our assertion is that with probability $1 - \eta$ the inequalities

$$\frac{1}{l}\sum_{i=1}^{l} t_i - \frac{\sigma}{\sqrt{l\eta}} < Mt < \frac{1}{l}\sum_{i=1}^{l} t_i + \frac{\sigma}{\sqrt{l\eta}} \tag{2.11}$$

are valid. (This assertion is completely equivalent to (2.10).)

If the variance σ^2 of the random variable t were known, the inequalities (2.11) would determine the size of the confidence interval for the mathematical expectation Mt and thus provide a *guaranteed estimator of the mean*, i.e., an estimator which is valid with a given probability. Therefore in order to obtain a guaranteed estimator of the mean based on the value of the empirical mean it is sufficient to know either an *absolute bound* τ^2_{abs} *on the variance*

$$\sigma^2 \leq \tau^2_{\text{abs}}, \tag{2.12}$$

or—provided the true mean value is a positive quantity—a bound on the *relative value of the variance*

$$\left(\frac{\sigma}{Mt}\right)^2 \leq \tau^2_{\text{rel}}. \tag{2.13}$$

Indeed, (2.11) and (2.12) imply that the knowledge of an absolute bound on the variance immediately leads to the construction of a guaranteed estimator of the form

$$\frac{1}{l}\sum_{i=1}^{l} t_i - \frac{\tau_{\text{abs}}}{\sqrt{\eta l}} < Mt < \frac{1}{l}\sum_{i=1}^{l} t_i + \frac{\tau_{\text{abs}}}{\sqrt{\eta l}}. \tag{2.14}$$

Also (2.11) and (2.13) imply that the knowledge of a bound on the relative variance leads for $l > \tau_{\text{rel}}^2/\eta$ to the construction of a guaranteed estimator of the form

$$\frac{\frac{1}{l}\sum_{i=1}^{l} t_i}{1 + \frac{\tau_{\text{rel}}}{\sqrt{\eta l}}} < Mt < \frac{\frac{1}{l}\sum_{i=1}^{l} t_i}{1 - \frac{\tau_{\text{rel}}}{\sqrt{\eta l}}}. \tag{2.15}$$

Now let the random variable t be nonnegative (this is the case studied in this book, since $t_\alpha = Q(z, \alpha) = (y - F(x, \alpha))^2$). Then *a fortiori* $Mt > 0$, and hence one can utilize information on the bound of the relative variance.

To obtain confidence intervals (2.14) and (2.15), Chebyshev's inequality was utilized. This inequality is valid for arbitrary distributions, with finite variances and therefore for some distributions it may be very coarse. In particular, if a distribution is such that the variable t is positive and is bounded by τ (in this case $\sigma \leq \tau/2$), then a more refined bound than Chebyshev's inequality is valid (Hoeffding inequality):

$$P\left\{\left|\frac{1}{l}\sum_{i=1}^{l} t_i - Mt\right| \geq \varkappa\right\} \leq 2e^{-2\varkappa^2 l/\tau^2}. \tag{2.16}$$

Using (2.16), a more precise guaranteed estimator of the value of the mathematical expectation may be derived.

In order to be able to utilize the inequality (2.16), we shall require, instead of prior knowledge of the absolute bound on the variance of a positive random variable, information about the absolute bound τ of the random variable t itself (when such a bound exists). Thus in order to be able to estimate the mean based on the value of an empirical mean, it is sufficient to know either the absolute bound τ on the random variable t or a bound τ_{rel} on the relative variance of the random variable t.

In this book we shall study the distribution of a collection of random variables

$$t_\alpha = Q(z, \alpha) = (y - F(x, \alpha))^2,$$

depending on parameter α, rather than a single random variable t. To obtain uniformly guaranteed estimators of the mean values of these variables a uniform characteristic of large deviations for the variables will be required.

A possible deviation on the set $t_\alpha = Q(z, \alpha)$ will be characterized by an absolute bound on the loss function

$$\tau_{\text{abs}} = \sup_{\alpha, z} Q(z, \alpha) \tag{2.17}$$

or by a bound on the relative variance

$$\tau_{\text{rel}} = \sup_{\alpha}\left[\frac{D\{Q(z,\alpha)\}}{(MQ(z,\alpha))^2}\right]^{1/2} = \sup_{\alpha}\sqrt{\frac{MQ^2(z,\alpha)}{(MQ(z,\alpha))^2} - 1.}\dagger \qquad (2.18)$$

Below it will be shown that if at least one of these characteristics of deviations (an absolute or relative bound) is known, then based on a random sample of fixed size l one can provide a guaranteed estimator of the value of the expected risk, and under some additional restriction the problem of minimizing the expected risk can be solved.

Remark. In this section we have used the Chebyshev's inequality for the second central moment. The above reasoning can be made on the basis of the Chebyshev's inequality for an absolute central moment of any order $p > 1$ (even if p is not integer). In this case the possible deviations are characterized by

$$\sup_{\alpha}\sqrt[p]{M\left|\frac{Q(z,\alpha)}{MQ(z,\alpha)} - 1\right|^p} = \tau_p.$$

§3 Prior Information in Problems of Estimating Dependences on the Basis of Empirical Data

Thus to obtain a guaranteed solution for the problem of minimizing the expected risk on the basis of a limited amount of empirical data, it is necessary to utilize prior information concerning possible large deviations of random variables $t_\alpha = Q(z,\alpha)$. The size of possible deviations is characterized by either an absolute bound on the loss (2.17) or a bound on the relative variance (2.18). How bothersome is it to obtain prior information about absolute or relative bounds for the three problems of estimating dependences discussed in this book: pattern recognition, regression estimation, and interpretation of results of indirect experiments?

A remarkable property of the pattern recognition problem is that the absolute value of the loss is bounded here by 1. Indeed, according to the formulation of the recognition problem, the loss function

$$Q(z,\alpha) = (\omega - F(x,\alpha))^2$$

is either 0 or 1. Thus the prior absolute bound on the value of the loss exists trivially in this case.

In problems of regression estimation or interpretation of indirect experiments the existence of an absolute bound on the value of the loss is far from trivial. More often than not, no such bound exists. This may happen even

† The symbol D is used here and below to denote the variance operator.

in the case of estimating linear regression. Indeed, the loss function in this case equals

$$Q(z, \alpha) = (y - F(x, \alpha))^2,$$

and if there are no restrictions on the value of parameters α, then one can find—in the class of linear functions $F(x, \alpha)$—a function such that the value of the loss may be arbitrarily large even if the variables y and x are bounded. Therefore for solving problems of regression estimation and interpretation of indirect experiments we shall utilize information about a bound on the relative variance of losses rather than an absolute bound on possible losses.

What then is the relation between this prior information and the prior information usually utilized in problems of estimating dependencies? Fix a function $F(x, \alpha^*)$ in $Q(z, \alpha) = (y - F(x, \alpha))^2$. Then the probability density $P(x, y)$ generates a random variable

$$t'_{\alpha^*} = y - F(x, \alpha^*),$$

and hence a bound on the relative variance is the prior information on the probability density of random variables $(t'_{\alpha^*})^2$.

If the distribution of t'_{α^*} is Gaussian for any α^*, then a bound on the relative variance of losses is equal to

$$\tau_{\text{rel}} = \sqrt{\frac{M(t'_\alpha)^4}{(M(t'_\alpha)^2)^2} - 1}$$

$$= \sqrt{\frac{\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(t'_\alpha)^4 \exp\left\{-\frac{(t'_\alpha - \mu)^2}{2\sigma^2}\right\} dt'_\alpha}{\left(\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(t'_\alpha)^2 \exp\left\{-\frac{(t'_\alpha - \mu)^2}{2\sigma^2}\right\} dt'_\alpha\right)^2} - 1} = \sqrt{2}$$

independently of the parameters of the distribution. If the distribution of t'_{α^*} is uniform for any α^*, then the bound is

$$\tau_{\text{rel}} = \sup_{a,b} \sqrt{\frac{\frac{1}{b-a}\int_a^b (t'_\alpha)^4 \, dt'_\alpha}{\left(\frac{1}{b-a}\int_a^b (t'_\alpha)^2 \, dt'_\alpha\right)^2} - 1} = \sqrt{\frac{4}{5}} = \sqrt{0.8}.$$

Finally, if the distribution of t_{α^*} for any α^* is Laplacian (double-exponential) then the bound is

$$\tau_{\text{rel}} = \sqrt{\frac{\frac{1}{2\Delta}\int_{-\infty}^{\infty}(t'_\alpha)^4 \exp\left\{-\left|\frac{t'_\alpha - \mu}{\Delta}\right|\right\} dt'_\alpha}{\left(\frac{1}{2\Delta}\int_{-\infty}^{\infty}(t'_\alpha)^2 \exp\left\{-\left|\frac{t'_\alpha - \mu}{\Delta}\right|\right\} dt'_\alpha\right)^2} - 1} = \sqrt{5}.$$

This bound does not depend on the parameters of the distributions.

Prior information on the distribution in terms of a bound on the relative variance of losses is the minimal prior information which is utilized in this book.

Another kind of prior information which is usually utilized for the estimation of functional dependence (see Chapters 3, 4, and 5) is the type of probability density of the random variable $t'_{\alpha^*} = y - F(x, \alpha^*)$ (for example, the Gaussian law or Laplacian law). The necessity of providing this prior information is a much stronger requirement than the provision of a bound on the relative variance of losses. Indeed the assumption that $\tau_{\text{rel}} < 2.5$ may be satisfied for Gaussian, uniform, Laplace, and many other distributions, while the assumption of a specific form of distribution allows us to obtain results which are guaranteed only for this particular type of distribution.

§4 Two Procedures for Minimizing the Expected Risk

In this section we shall assume that an absolute bound on the value of possible losses is given:

$$\sup_{z,\alpha} Q(z, \alpha) = \tau_{\text{abs}}.$$

Our purpose, based on a random independent sample

$$z_1, \ldots, z_l, \tag{2.19}$$

is to construct an empirical functional

$$I_{\text{emp}}(\alpha) = \Phi(Q(z_1, \alpha), \ldots, Q(z_l, \alpha); z_1, \ldots, z_l),$$

whose minimum point $\alpha = \alpha^*$ yields (with a given probability $1 - \eta$) a value for the expected-risk functional

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz \tag{2.20}$$

close to the minimal one.

There is a "natural" method for constructing such a functional. One estimates, from the sample (2.19), the probability density $\hat{P}(z)$, and then substitutes into (2.20) the estimated density $\hat{P}(z)$ in place of $P(z)$. The functional obtained does not depend on the unknown density and at least in principle may be minimized.

It seems that the problem of minimizing the expected risk on the basis of empirical data is reduced to an estimation of the probability density. In turn, the problem of estimating the probability density from a random independent sample is a central problem of mathematical statistics. Thus a solution to a particular problem of statistics, the minimization of the expected

risk on the basis of empirical data, depends on a solution to its central problem.

In the next section we shall discuss in detail the formulation of the problem of the estimation of a density; in this section we shall establish two distinct procedures which allow us to solve the problem of minimizing the expected risk on the basis of empirical data. One of these procedures is indeed based on the fact that the estimated density $\hat{P}(z)$ approaches the actual one, while the other procedure has a completely different theoretical basis.

Thus let $0 \leq Q(z, \alpha) \leq \tau$. Consider two types of empirical functionals: one of the type

$$I'_{\text{emp}}(\alpha) = \int Q(z, \alpha)\hat{P}(z)\, dz, \tag{2.21}$$

where $\hat{P}(z)$ is an empirical density estimated from the sample $z_1, \ldots, z_l$, and the other of the type

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha). \tag{2.22}$$

The functional (2.22) is usually called a *functional of empirical risk.*

Formally a functional of empirical risk is a particular case of (2.21). Indeed, if for the approximating density in (2.21) one chooses the density

$$\hat{\hat{P}}_\varepsilon(z) = \frac{1}{l} \sum_{i=1}^{l} \pi_\varepsilon(z - z_i), \tag{2.23}$$

where, for example

$$\pi_\varepsilon(z) = \frac{1}{(\sqrt{2\pi}\,\varepsilon)^n} \exp\left\{-\frac{z^T z}{2\varepsilon^2}\right\}$$

(n is the dimension of the vector z), then as $\varepsilon \to 0$ it can be shown that $I'_{\text{emp}}(\alpha) \to I_{\text{emp}}(\alpha)$. (Here we utilize the relationship $\lim_{\varepsilon \to 0} \pi_\varepsilon(z) = \delta(z)$.) However, it makes sense to single out the functional (2.22), since the success of minimizing the expected risk by minimizing (2.21) and by minimizing (2.22) is determined by different factors. In the first case the success is due to the proximity between the estimated density and the actual one, while in the second case the density $\hat{\hat{P}}_\varepsilon(z)$ for small ε does not approach $P(z)$. Nevertheless under certain conditions the minimum point for a functional of empirical risk yields a value of the functional (2.20) which is close to the minimal.

Indeed, let $\hat{P}(z)$ be close to $P(z)$, i.e.,

$$\int |P(z) - \hat{P}(z)|\, dz \leq \varepsilon,$$

and let the minimum of the empirical functional be attained at $\alpha = \alpha_{\text{emp}}$, while the minimum of the expected risk is attained at $\alpha = \alpha_0$. Then the

following chain of inequalities is valid

$$I(\alpha_{\text{emp}}) - I(\alpha_0) \leq I(\alpha_{\text{emp}}) - I'_{\text{emp}}(\alpha_{\text{emp}}) + I'_{\text{emp}}(\alpha_0) - I(\alpha_0)$$
$$\leq \int Q(z, \alpha_{\text{emp}}) |P(z) - \hat{P}(z)| dz + \int Q(z, \alpha_0) |P(z) - \hat{P}(z)| dz$$
$$\leq 2\tau_{\text{abs}} \varepsilon,$$

which implies the proximity between the minima of the functionals (2.20) and (2.21).

We now show that the approximating density (2.23) does not approach the actual one as $\varepsilon \to 0$. Let $P(z)$ be a bounded function. Subdivide the set Z into two subsets: a set $\overline{Z}$ of a small measure containing all the sample elements, and the complementary set $Z \backslash \overline{Z}$.

It is easy to verify that for ε sufficiently small a set $\overline{Z}$ can be chosen so that

$$\int_Z |P(z) - \hat{\hat{P}}_\varepsilon(z)| dz \approx \int_{Z \backslash \overline{Z}} P(z) dz + \int_{\overline{Z}} \hat{\hat{P}}_\varepsilon(z) dz \simeq 2.$$

Thus success in minimizing the expected risk (2.20) using the method of minimizing a functional of empirical risk (2.22) is determined not by proximity between densities but by some other mechanism. Below in Section 6 it will be shown that this mechanism is based on the property of uniform convergence of empirical means to mathematical expectations over some set of events.

§5 The Problem of Estimating the Probability Density

Problems which are solved in probability theory on the one hand and mathematical statistics on the other are interrelated as direct and inverse.

Problems in probability theory can be described by the following setup: the composition of a general population and the probability distribution law are known. It is required for a given scheme of experiments to estimate the probabilities of outcomes of the experiment.

Mathematical statistics solves inverse problems: based on the results of an experiment, it is required to determine properties of the distribution law. An "exhaustive" characteristic of a distribution law is the probability density (if the latter exists).

Thus the problem of estimating the probability density from a sample is a central problem of mathematical statistics. In this section we shall verify that the problem of density estimation is usually an ill-posed one.

Let a sample $t_1, \ldots, t_l$ be given, and a class of functions to which the probability density $P(t)$ belongs to broadly defined (i.e., it is known only that $P(t)$ belongs to continuous functions). It is required to estimate the probability density.

First consider the one-dimensional case. By definition the probability density $P(t)$ is related to the cumulative distribution function $F(z) = P\{t \leq z\}$ as follows:

$$\int_{-\infty}^{z} P(t)\,dt = F(z),$$

or equivalently

$$\int_{-\infty}^{\infty} \theta(z - t)P(t)\,dt = F(z), \tag{2.24}$$

where

$$\theta(x) = \begin{cases} 1 & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

For continuous densities there is a unique solution of the integral equation (2.24).

Now define an empirical cumulative distribution function: $F_l(z) = k/l$ if z exceeds k terms of the sample $z_1, \ldots, z_l$. The basic theorem of mathematical statistics—the Glivenko–Cantelli theorem—asserts that as the sample size l increases, the empirical cumulative distribution function uniformly approaches the actual one.

Theorem (Glivenko–Cantelli). *Let $F(z)$ be a cumulative distribution function of a random variable z, and $F_l(z)$ be the empirical cumulative distribution function. Then*

$$P\left\{\sup_z |F(z) - F_l(z)| \xrightarrow[l\to\infty]{} 0\right\} = 1.$$

We shall not prove this theorem here. In Chapter 6 a theorem on uniform convergence of relative frequencies of occurences of events to their probabilities is proved. The Glivenko–Cantelli theorem follows from it as a particular case.

We now return to the integral equation (2.24) whose solution determines the probability density. We seek an approximate solution of this equation in those situations when instead of a cumulative distribution function $F(z)$ an empirical cumulative distribution function $F_l(z)$ is known from a finite sample. In Chapter 9, utilizing a bound on the rate of uniform convergence of $F_l(z)$ to $F(z)$, we shall show that there exists a procedure for obtaining approximate solutions of Equation (2.24) such that as l increases the sequence of solutions tends to the required probability density.

Thus it is possible in principle to estimate a continuous probability density. However, estimating a density is associated with the solution of the ill-posed problem of numerical differentiation of (2.24) under conditions where the right-hand side of the equation is imprecisely defined.

Actually, in the case of estimating a probability density it is known *a priori* that a solution for the integral equation (2.24) is not an arbitrary continuous function but rather a function $P(t)$ which takes on nonnegative values only and satisfies the condition

$$\int_{-\infty}^{\infty} P(t)\,dt = 1.$$

However, this prior information is not sufficient for a problem of solving integral equation (2.24) to become well posed.

Analogously to the one-dimensional case, the problem of estimating a multidimensional density can be posed. For this purpose we write the integral equation which connects a multidimensional density with a multidimensional cumulative distribution function:

$$\int_{-\infty}^{z^1} \cdots \int_{-\infty}^{z^n} P(t^1, \ldots, t^n)\,dt^1 \cdots dt^n = P(t^1 \leq z^1; \ldots; t^n \leq z^n), \tag{2.25}$$

and define a multidimensional empirical cumulative distribution function by

$$F_l(z^1, \ldots, z^n) = \frac{k}{l}, \tag{2.26}$$

where k is the number of elements of the sample $z_1, \ldots, z_l$ which fall into the region $t^1 \leq z^1, \ldots, t^n \leq z^n$.

It turns out that a multivariate analog of the Glivenko–Cantelli theorem is valid: as the sample size increases the empirical cumulative distribution function converges uniformly to the population cumulative distribution function. The validity of the generalized Glivenko–Cantelli theorem also follows from the general theory of uniform convergence of frequencies to the corresponding probabilities discussed in Chapter 6. Using this theorem analogously to the one-dimensional case, one establishes the possibility—in principle—of estimating the multidimensional density from empirical data.

Thus the problem of estimating the density in the class of continuous functions is reduced to an ill-posed problem of numerical differentiation of a cumulative distribution function.†

Observe that the formulation of the problem of numerical differentiation presented here differs from the problem of numerical differentiation considered in Example 3 of Chapter 1. There an ill-posed measurement problem was considered, i.e., formulations of ill-posed problems for which the errors were results of measurements (observations) and the values of the right-hand side of the integral equation (2.24) were defined statistically independently at l points. In the present case the difference between the exact value of the

† There are nonparametric methods for estimating the density (e.g., Parzen's method) which seem to avoid the necessity of solving ill-posed problems. However, as will be shown in Chapter 9, problems which arise in the actual realization of these methods are equivalent to ill-posed problems of numerical differentiation.

right-hand side and the function obtained as a result of observations is a random function.

Thus the problem of estimating the probability density is more general than the interpretation of results of indirect experiments. Hence it would seem unreasonable to solve the problem of minimizing the expected risk on the basis of empirical data by means of estimating a probability density. (Quite the reverse, in Chapter 9 we shall consider the problem of density estimation as a problem of minimizing expected risk based on empirical data.)

However, some degenerate cases are possible where there is available substantial prior information about the density to be estimated, so that the problem ceases to be ill posed. For example, the problem of density estimation may turn out to be well posed if the density is known up to a finite number of parameters. Methods of estimation of a density defined up to a finite number of parameters are called *methods of parametric statistics.* They form a special class of methods which are significantly different from the general methods of density estimation. (The latter are sometimes called *methods of non-parametric statistics.*)

§6 Uniform Proximity Between Empirical Means and Mathematical Expectations

Above it was established that there exist two procedures for minimizing the expected risk on the basis of empirical data.

The first is connected with minimization of an empirical functional constructed from the estimated density. However, the intermediate problem—the density estimation—is in general more complex than the problem of risk minimization based on empirical data. Therefore it is generally unreasonable to solve the problem of minimizing the expected risk by means of density estimation.

Here we shall consider the second procedure. We shall minimize the expected risk

$$I(\alpha) = \int Q(z, \alpha)P(z)\,dz$$

on the basis of the data

$$z_1, \ldots, z_l$$

by minimizing the functional of the empirical risk,

$$I_{\text{emp}}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} Q(z_i, \alpha).$$

For each fixed $\alpha = \alpha^*$ the functional $I(\alpha^*)$ determines the mathematical expectation for a random variable $t_{\alpha^*} = Q(z, \alpha^*)$, while the functional $I_{\text{emp}}(\alpha^*)$ is the empirical (arithmetic) mean of this random variable.

According to the classical theorems of probability theory, in sufficiently general cases the empirical mean of the random variable t_{α^*} converges as l increases to the mathematical expectation of this random variable. However, these theorems do not imply that the value of parameter α_{emp} which yields the minimum of the empirical risk $I_{\text{emp}}(\alpha)$ will also yield a value of the expected risk $I(\alpha)$ which is close to the minimal one. This is an important assertion, and we shall discuss it in greater detail.

Assume for concreteness that the parameter α is a scalar in the interval $[0, 1]$. A value $I(\alpha)$ corresponds to each α. Consider the function $I(\alpha)$. Along with this function consider the function $I_{\text{emp}}(\alpha)$ which for each α determines the empirical mean obtained on the basis of a sample of size l (Figure 2).

The method of minimizing empirical risk proposes to decide about the minimum of the function $I(\alpha)$ on the basis of the minimum of the function $I_{\text{emp}}(\alpha)$. In order to be able to do this it is sufficient that the curve $I_{\text{emp}}(\alpha)$ be located entirely within a $\varkappa$-tube of the curve $I(\alpha)$. A large deviation at even one point (as in Figure 2) may result in a point of large deviation of $I_{\text{emp}}(\alpha)$ being chosen as the minimizing point of $I(\alpha)$. In this case the minimum of $I_{\text{emp}}(\alpha)$ does not in any way characterize the minimum of $I(\alpha)$. If, however, the function $I_{\text{emp}}(\alpha)$ approximates $I(\alpha)$ uniformly in α with precision $\varkappa$, then the minimum of $I_{\text{emp}}(\alpha)$ deviates from the minimum of $I(\alpha)$ by an amount not exceeding $2\varkappa$. Formally this means that we are interested not in the classical condition that for any α and κ the relation

$$P\{|I(\alpha) - I_{\text{emp}}(\alpha)| > \varkappa\} \underset{l \to \infty}{\longrightarrow} 0 \tag{2.27}$$

is valid, but in a more stringent condition that for any $\varkappa$ the relation

$$P\left\{\sup_{\alpha} |I(\alpha) - I_{\text{emp}}(\alpha)| > \varkappa\right\} \underset{l \to \infty}{\longrightarrow} 0 \tag{2.28}$$

holds. When (2.28) is satisfied we say that a *uniform convergence in the parameter α of empirical means to their mathematical expectation* occurs.

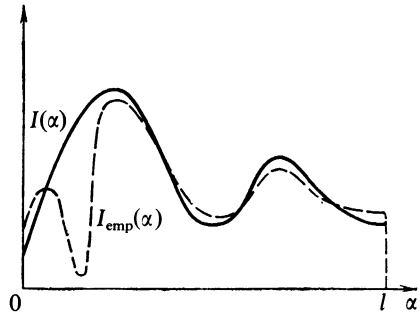

Figure 2

The second procedure for minimizing risk is connected with uniform convergence in parameter α of empirical means to their mathematical expectations. However, for our purposes—the minimization of expected risk based on a sample of fixed size—the simple fact of uniform convergence is not sufficient. In order to be able, with a given probability, to guarantee obtaining solutions which yield a value of the functional close to the minimal one, it is necessary to know a bound on the rate of uniform convergence. Indeed the fulfillment of the inequality

$$P\left\{\sup_{\alpha} |I(\alpha) - I_{\text{emp}}(\alpha)| \geq \varkappa\right\} < \eta(l, \varkappa),$$

$$\lim_{l \to \infty} \eta(l, \varkappa) = 0$$

is equivalent to the following assertion: with probability $1 - \eta(l, \varkappa)$ the bound

$$I_{\text{emp}}(\alpha) - \varkappa \leq I(\alpha) \leq I_{\text{emp}}(\alpha) + \varkappa \tag{2.29}$$

is valid simultaneously for all α. If, however $\eta(l, \varkappa)$ is a decreasing function in l and $\varkappa$, then for the given confidence level $1 - \eta$,

$$\eta(l, \varkappa) = \eta, \tag{2.30}$$

the size of the confidence interval $\varkappa = \varkappa(l, \eta)$ obtained as the solution of Equation (2.30) decreases with increasing l. Consequently for l large the point α_{emp} of the minimum of empirical risk will yield a value of the expected risk close to the minimal one. For any fixed l one can assert that with probability $1 - \eta$ the point α_{emp} yields a value of the expected risk belonging to the interval

$$I_{\text{emp}}(\alpha_{\text{emp}}) - \varkappa \leq I(\alpha_{\text{emp}}) \leq I_{\text{emp}}(\alpha_{\text{emp}}) + \varkappa.$$

§7 A Generalization of the Glivenko–Cantelli Theorem and the Problem of Pattern Recognition

In this section we shall consider the particular case where the loss function $Q(z, \alpha)$ of the functional

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz \tag{2.31}$$

admits only the two values, 0 and 1. As we already noted, the problem of pattern recognition reduces to this case.

Denote by $S(\alpha^*)$ the set of vectors z for which the given loss function $Q(z, \alpha^*)$ admits the value 1. In other words $S(\alpha^*)$ is the event $S(\alpha^*) = \{z : Q(z, \alpha^*) = 1\}$. For a fixed

$\alpha = \alpha^*$ the functional (2.31) determines the probability that the vector z belongs to the set $S(\alpha^*)$, i.e., the probability of the event $S(\alpha^*)$.

Correspondingly, for each fixed $\alpha = \alpha^*$ the functional of empirical risk

$$I_{\text{emp}}(\alpha^*) = \frac{1}{l}\sum_{i=1}^{l} Q(z_i, \alpha^*) \tag{2.32}$$

determines the frequency of the event $S(\alpha^*)$ obtained from the sample $z_1, \ldots, z_l$ of size l. In order to single out this important particular case we shall denote the functional (2.31) by $P(\alpha)$ and the functional (2.32) by $\nu(\alpha)$. In this notation the condition (2.28) will be written in the form

$$P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\} \underset{l\to\infty}{\to} 0.$$

This indicates a uniform convergence of frequencies of occurrence of events to their probabilities over the class of events $S(\alpha)$. In these terms the assertion of the Glivenko–Cantelli theorem that the empirical cumulative distribution function uniformly converges to the population cumulative distribution function is an assertion about the existence of the uniform convergence of frequencies of events to their probabilities for a special system of events.

Indeed, consider the line z and a set of rays $z \leq \alpha$. This set of rays defines a system of events $S^1(\alpha)$ (the event $S^1(\alpha^*)$ is that the point z belongs to the ray $z \leq \alpha^*$). In these terms the assertion of the Glivenko–Cantelli theorem is as follows: "a uniform convergence of frequencies of events to their probabilities is valid over a class of events $S^1(\alpha)$".

Consider now the following class of events $S^n(\alpha)$: a vector $z = (z^1, \ldots, z^n)^T$ belongs to the event $S^n(\alpha^*)$ (here $\alpha^* = \alpha_1^*, \ldots, \alpha_n^*)^T$) if simultaneously for all n coordinates the inequalities $z^1 \leq \alpha_1^*, \ldots, z^n \leq \alpha_n^*$ are fulfilled. The set of all events $S^n(\alpha^*)$ is the class $S^n(\alpha)$. In these terms the multivariate analog of the Glivenko–Cantelli lemma is the assertion of uniform convergence of frequencies of occurrences of events to their probabilities over the class of events $S^n(\alpha)$.

Thus the condition of uniform convergence of frequencies of occurrences of events to their probabilities for various systems of events which occurs in the study of the pattern recognition problem leads to a generalization of the Glivenko–Cantelli lemma.

§8 Remarks on Two Procedures for Minimizing Expected Risk on the Basis of Empirical Data

Thus there exist two methods for minimizing the expected risk on the basis of empirical data. One is connected with the feasibility of estimating the probability density, and the other with the possibility of assuring a uniform convergence of empirical means to their mathematical expectations.

It makes sense to estimate the density only in the trivial case when substantial prior information is given. If the prior information is limited, then the solution of the intermediate problem—estimation of the density—turns out to be no simpler than the problem of minimizing the expected

risk. In this case the possibility of density estimation is based on the Glivenko–Cantelli theorem, i.e., on the existence of uniform convergence of frequencies of events to their respective probabilities for a special class of events.

The second procedure for risk minimization is directly based on the existence of uniform convergence of empirical means to their mathematical expectations.

Below, in Chapters 6–7, it will be shown that sufficient conditions for the existence of uniform convergence of the means to mathematical expectations are determined by special features of the loss functions. For the problem of estimating dependences the requirement is that the class in which estimation is carried out should be a rather narrow one.

The existence of two procedures of minimizing the expected risk reflects the presence of conditions of two types under which minimization of the expected risk based on empirical data is feasible in principle. Conditions of the first type connect the feasibility of risk minimization with the information available about the class of densities to which the estimated density belongs. In those cases when the density can be estimated one can successfully minimize the expected risk, regardless of the loss function (provided it does not admit large deviations). Conditions of the second type impose restrictions on the properties of loss functions and then independently of the structure of $P(z)$ so that it is possible to successfully minimize the expected risk.

When solving problems of estimating dependences on the basis of empirical data under the condition that the loss function does not admit large deviations, the difference between these two approaches is reflected in the set-ups of possible assertions:

Assertions of the first type. If the nature of the problem is well diagnosed (a "narrow" class of densities $\{P(z)\}$ to which the required density belongs is found), then independently of the special features of the class of functions in which the estimation takes place, the minimum of the empirical functional will be close to the minimum of the expected risk.

Assertions of the second type. If the estimation is taking place in a sufficiently "narrow" class of functions $F(x, \alpha)$, then regardless of the nature of the problem (i.e., the density $P(z)$), the minimum of the empirical risk will be close to the minimum of the expected risk.

It should be noted that formally there is a certain advantage to utilizing algorithms for which assertions of the second type are feasible. Indeed, assertions of the first type require that:

(1) the class densities in which estimation is carried out be sufficiently narrow, and
(2) the required density belong to this class.

Assertions of the second type involve only one requirement: that the class of functions in which estimation takes place be sufficiently narrow. In practice,

it is not difficult to control the width of the class of densities as well as that of the functions.

The problem of whether the estimated density belongs to a given class is always open.

The main thrust of this book is to determine conditions of uniform convergence and utilize these for the estimation of dependences based on samples of a fixed size. Utilizing bounds on the rate of uniform convergence of means to their mathematical expectations, it becomes possible not only to establish the method of minimizing empirical risk, but also to construct a new method for minimizing risk (the method of structural minimization) which allows us under conditions of limited empirical data to arrive at a solution which yields the smallest guaranteed value of the expected risk.

Chapters 6–10 are devoted to a description of the methods of risk minimization utilizing procedures of uniform convergence. However, before studying this procedure systematically we shall consider classical methods of risk minimization based on the idea of minimizing a functional constructed by means of the estimated density. As was mentioned above, in exceptional cases (when the density is known up to a finite number of parameters) the estimation problem may be stable and its solution—as well as that of estimation of dependences from empirical data—may be successfully achieved using methods of parametric statistics. In Chapter 3 the application of parametric statistics to solutions of problems of pattern recognition is discussed, and in Chapters 4 and 5 these methods are applied to regression estimation.

Chapter 3

Methods of Parametric Statistics for the Pattern Recognition Problem

§1 The Pattern Recognition Problem

It is required to minimize the functional

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \tag{3.1}$$

under the conditions when the density $P(x, y)$ is unknown but the sample

$$x_1, y_1; \ldots; x_l, y_l \tag{3.2}$$

is given, based on random independent trials according to $P(x, y)$.

We shall solve this problem applying the following scheme:

(1) Estimate the density from the sample (3.2). Denote the estimated density by $\hat{P}(x, y)$.
(2) Construct the functional

$$I_{\text{emp}}(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(x, y)\, dx\, dy \tag{3.3}$$

using the estimated density.
(3) Obtain the minimum of this functional, and declare the function $F(x, \alpha_{\text{emp}})$ which yields the minimum of (3.3) to be the solution of the original minimization problem (3.1).

As was pointed out in Chapter 2, this scheme can be successfully carried out only if substantial prior information concerning the density $P(x, y)$ is available (namely, when the density is completely specified up to its parameters). In other words, success can be achieved if the model of the

estimated density is known. The model of the required density turns out to be quite different for different problems of estimating dependences.

In this chapter we shall consider the pattern recognition problem. A characteristic feature of this problem is that the unknown probability density† $P(x, \omega)$ can be represented as a union of two densities $P(x|\omega = 0)$ and $P(x|\omega = 1)$ defined on different subspaces $X, 0$ and $X, 1$:

$$P(x, \omega) = P(x|\omega = 0)P(\omega = 0)(1 - \omega) + P(x|\omega = 1)P(\omega = 1)\omega. \tag{3.4}$$

The set of pairs x, ω consists of two nonoverlapping subspaces of dimensionality n, namely

$$X \subset E_n, \quad \omega = 0 \quad \text{and} \quad X \subset E_n, \quad \omega = 1.$$

The formula (3.4) asserts that on the first subspace the density is equal to $P(x|\omega = 0)P(\omega = 0)$, and on the second $P(x|\omega = 1)P(\omega = 1)$. In formula (34) $P(x|\omega = 0)$ and $P(x|\omega = 1)$ are the components of the union; $P(\omega = 0)$ and $P(\omega = 1) = 1 - P(\omega = 0)$ are the proportions.

Let the density $P(x, \omega)$ be known up to a finite number $m_1 + m_2 + 1$ of parameters

$$P(x, \omega) = P_\beta(x|\omega = 0)P(\omega = 0)(1 - \omega) + P_\gamma(x|\omega = 1)P(\omega = 1)\omega, \tag{3.5}$$

where β is an unknown m_1-dimensional vector of parameters of density $P_\beta(x|\omega = 0)$, γ is an unknown m_2-dimensional vector of parameters of the density $P_\gamma(x|\omega = 1)$, and $P(\omega = 0)$ is a scalar parameter.

Now in order to implement our scheme it is necessary to be able to solve two problems:

(1) to find the minimum of functional (3.3) for a given density $P(x, \omega)$;
(2) based on the sample (3.2), to estimate the density of $P(x, \omega)$.

The first problem is referred to in statistics as the *problem of discriminant analysis*; the second is called the *problem of estimating the density in a parametric class of functions*. We now consider these two problems.

§2 Discriminant Analysis

It is required to obtain the minimum of the functional (3.3) for a given density (given components of union $P(x|\omega = 0)$, $P(x|\omega = 1)$ and proportions $P(\omega = 0)$, $P(\omega = 1) = 1 - P(\omega = 0)$).

First consider the simple case: the class of possible decision rules $F(x, \alpha)$ is in no way restricted. In this situation it is easy to construct a minimizing

† We use the letter ω instead of y to emphasize that it takes only the two values 0 and 1.

rule which minimizes the functional (3.3). Indeed, according to Bayes's formula the probability that the vector x belongs to the first (second) class is determined by

$$P(\omega = 0|x) = \frac{P(x|\omega = 0)P(\omega = 0)}{P(x|\omega = 0)P(\omega = 0) + P(x|\omega = 1)P(\omega = 1)}$$
$$\left(P(\omega = 1|x) = \frac{P(x|\omega = 1)(1 - P(\omega = 0))}{P(x|\omega = 0)P(\omega = 0) + P(x|\omega = 1)P(\omega = 1)}\right). \tag{3.6}$$

Minimal loss (the minimum probability of error) can be obtained for the classification in which the vector x is assigned to the first class if its affiliation to the first class is more probable than to the second, i.e., if

$$P(\omega = 0|x) > P(\omega = 1|x).$$

Otherwise the vector x is assigned to the second class. In other words, taking (3.6) into account, the vector x should be assigned to the first class provided the inequality

$$\frac{P(x|\omega = 1)}{P(x|\omega = 0)} < \frac{P(\omega = 0)}{1 - P(\omega = 0)},$$

is fulfilled, or equivalently, the optimal classification of vectors is carried out by means of the indicator function

$$F(x) = \theta\left[\ln P(x|\omega = 1) - \ln P(x|\omega = 0) + \ln \frac{1 - P(\omega = 0)}{P(\omega = 0)}\right], \tag{3.7}$$

where

$$\theta(z) = \begin{cases} 1 & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

Therefore the knowledge of the probability density (composition and proportion of the union (3.5)) allows us to construct an optimal decision rule immediately.

However, the problem of finding an optimal decision rule becomes substantially more complex if the class of admissible decision rules $F(x, \alpha)$ is restricted. In particular, the problem of finding an optimal linear decision rule of the form

$$F(x, \alpha) = \theta[\alpha^T x + \alpha_0] \tag{3.8}$$

is a difficult one. The vector $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ determines the direction of a linear discriminant function, and the parameter α_0 its threshold value. The problem of finding the minimum of (3.3) in the class (3.8) is called the problem of *linear discriminant analysis.*

In the thirties R. A. Fisher proposed as the direction of the linear discriminant function a direction along which the maximum of the relative distance between the mathematical expectations of projections of vectors of different classes is obtained, i.e., the direction α along which the maximum of

$$T(\alpha) = \frac{(m_1(\alpha) - m_2(\alpha))^2}{\sigma_1^2(\alpha) + \sigma_2^2(\alpha)}, \tag{3.9}$$

where

$$m_1(\alpha) = \int \alpha^T x P_\beta(x|\omega = 0)\, dx,$$

$$m_2(\alpha) = \int \alpha^T x P_\gamma(x|\omega = 1)\, dx,$$

$$\sigma_1^2(\alpha) = \int (\alpha^T x - m_1(\alpha))^2 P_\beta(x|\omega = 0)\, dx,$$

$$\sigma_2^2(\alpha) = \int (\alpha^T x - m_2(\alpha))^2 P_\gamma(x|\omega = 1)\, dx,$$

$$\alpha^T \alpha = 1$$

is attained.

The determination of the maximum of (3.9) for arbitrary densities is a very difficult problem. Therefore basic investigations in the area of linear discriminant analysis were directed first toward verifying for specific types of densities that Fisher's linear discriminant function indeed determines a solution of linear discriminant analysis, and secondly toward finding algorithms for computing the discriminant function. The basic result was that for the union of two normal laws

$$P(x|\omega = 0) = N(\mu_1, \Delta_1), \qquad P(x|\omega = 1) = N(\mu_2, \Delta_2)$$

(μ_1 is the mean vector, Δ_1 is the covariance matrix for the first multivariate normal distribution, and μ_2, Δ_2 are the analogous parameters for the second distribution), taken in proportions $P(\omega = 0)$ and $1 - P(\omega = 0)$, the optimal linear discriminant function is given by the direction vector

$$\alpha_{t^*} = (\mu_1 - \mu_2)^T (t^* \Delta_1 + (1 - t^*)\Delta_2)^{-1}, \tag{3.10}$$

where $0 \leq t^* \leq 1$. The value t^* is determined as the root of the so-called *resolvent function*

$$f(t) = t\sigma_1^2(\alpha_t) + (1 - t)\sigma_2^2(\alpha_t) - \ln\left(\frac{P(\omega = 0)}{1 - P(\omega = 0)} \cdot \frac{\sigma_2^2(\alpha_t)}{\sigma_1^2(\alpha_t)}\right). \tag{3.11}$$

For $P(\omega = 0) = \frac{1}{2}$ the direction (3.10) of the linear discriminant function maximizes the functional

$$I(\alpha) = \frac{(m_1(\alpha) - m_2(\alpha))^2}{t^*\sigma_1^2(\alpha) + (1 - t^*)\sigma_2^2(\alpha)}.$$

The calculation of the roots of the resolvent equation (3.11) is quite a difficult task. Therefore in practice when constructing a linear discriminant function it is assumed that $t^* = \frac{1}{2}$, and Fisher's linear discriminant is taken to be the solution of the problem. (More details are given in [71].)

Thus problems arising in discriminant analysis are due to the fact that the class of possible decision rules on which the minimum of functional (3.3) is to be determined is bounded. Therefore it may seem that the problem of discriminant analysis is artificial. Indeed, if it is possible to estimate probability density, what is the need for seeking a decision rule which yields a conditional minimum of the functional, when it is easy to find a decision rule (cf. (3.7)) which yields an absolute minimum for the functional (3.3)?

The fact of the matter is that if the density is estimated imprecisely, then the value of the guaranteed deviation of the minimum for the empirical functional from the minimum for the expected risk functional becomes larger for a function chosen from a wider class. Therefore it may happen that the smaller value of the guaranteed expected risk will be achieved, not at a function yielding the absolute minimum for the empirical functional, but rather on a function belonging to a narrower class and yielding the conditional minimum.

This result is connected with the effect of the second procedure for minimizing the expected risk (cf. Chapter 2, Section 4). The idea of narrowing the class of decision rules in order to obtain a smaller guaranteed value of the expected risk will be implemented below in Chapters 8 and 9. In the present chapter we shall consider parametric methods of estimating densities. In view of (3.7), the knowledge of the density immediately leads to the construction of a decision rule yielding the absolute minimum for (3.3).

§3 Decision Rules in Problems of Pattern Recognition

Algorithms of pattern recognition based on estimation of the density (gives components of the union (mixture) $P(x|\omega = 0)$ and $P(x|\omega = 1)$ and its proportion $P(\omega = 0)$) are traditionally associated with two classes of distributions.

3.1 First Class of Distributions

The probability distribution $P_\omega(x) = P(x|\omega)$ is such that coordinates of the vector $x = (x^1, \ldots, x^n)^T$ are statistically independent, i.e.,

$$P_\omega(x) = P_\omega(x^1) \cdots P_\omega(x^n), \qquad \omega = 0, 1, \tag{3.12}$$

and moreover each coordinate x^i of the vector x can take on only a fixed number of values. Let us assume that each coordinate x^i takes on τ_i values $c^i(1), \ldots, c^i(\tau_i)$. Thus in the case under consideration the distribution laws of random variables $P_{\omega=0}(x)$ and $P_{\omega=1}(x)$ are defined by the expression

(3.12), where $P_\omega(x^i)$ can be written as

$$P_\omega(x^i) = \begin{cases} p^i_\omega(1) & \text{for } x^i = c^i(1), \\ \vdots & \\ p^i_\omega(\tau_i) & \text{for } x^i = c^i(\tau_i), \end{cases} \tag{3.13}$$

$$\sum_{j=1}^{\tau_i} p^i_\omega(j) = 1.$$

Here $p^i_\omega(j)$ is the probability that for a vector belonging to the class $\omega = \{0, 1\}$ the value of the x^ith coordinate equals $c^i(j)$. To estimate the probability distribution for such a union means to find values of

$$r = 2\sum_{i=1}^{n} \tau_i + 1$$

parameters ($\sum_{i=1}^{n} \tau_i$ parameters for estimating each distribution $P_\omega(x)$, and one parameter—the proportion of the union).

According to (3.7) an optimal decision rule for the mixture formed by the two distributions (3.12) will be the following linear discriminant function:

$$F(x) = \theta\left(\sum_{i=1}^{n} \ln \frac{P_{\omega=1}(x^i)}{P_{\omega=0}(x^i)} - \ln \frac{p}{1-p}\right),$$

where p, $1 - p$ are proportions of the union.

3.2 Second Class of Distributions

Here in each class $\omega = \{0, 1\}$ vectors x are distributed according to the multivariate normal distribution

$$P_\omega(x) = \frac{1}{(2\pi)^{n/2}|\Delta_\omega|^{1/2}} \exp\{-\tfrac{1}{2}(x - \mu_\omega)^{\mathrm{T}} \Delta_\omega^{-1}(x - \mu_\omega)\},$$

where μ_ω is the vector of mean values and Δ_ω is the covariance matrix.

It follows from (3.7) that the optimal decision rule in this case becomes the quadratic discriminant function

$$F(x) = \theta[\tfrac{1}{2}(x - \mu_0)^{\mathrm{T}} \Delta_0^{-1}(x - \mu_0) - \tfrac{1}{2}(x - \mu_1)^{\mathrm{T}} \Delta_1^{-1}(x - \mu_1) + \ln \frac{|\Delta_0|}{|\Delta_1|} - \ln \frac{p}{1-p}], \tag{3.14}$$

where $\mu_0, \Delta_0; \mu_1, \Delta_1$ are parameters of the normal distributions forming the union (3.5) and p, $1 - p$ are the corresponding proportions. In the particular case when $\Delta_0 = \Delta_1 = \Delta$ the quadratic discriminant function (3.14) reduces to a linear one:

$$F(x) = \theta\left[(\mu_1 - \mu_0)^{\mathrm{T}} \Delta^{-1} x + \tfrac{1}{2}(\mu_0^{\mathrm{T}} \Delta^{-1} \mu_0 - \mu_1^{\mathrm{T}} \Delta^{-1} \mu_1) - \ln \frac{p}{1-p}\right].$$

§4 Evaluation of Qualities of Algorithms for Density Estimation

Thus the construction of a discriminant function based on empirical data reduces to an estimation of the probability distributions $P(x|\omega = 0)$ and $P(x|\omega = 1)$ and of the parameter p. The parameter p determines the fraction of pairs x, ω with $\omega = 0$ and may be estimated by the quantity $\tilde{p} = m/l$, where m is the number of pairs in the sample with $\omega = 0$ and l is the sample size.†

What are the algorithms that one should utilize for estimating the probability densities $P(x|\omega = 0)$, $P(x|\omega = 1)$? To answer this question one should first agree on the method of assessing qualities of estimating algorithms on the basis of samples of fixed size.

The quality of specific algorithm A which estimates the density $P(x, \alpha_0)$ from a sample $x_1, \ldots, x_l$ is naturally defined as the distance between the density and the estimated function $P_A(x|x_1, \ldots, x_l)$, i.e., by the quantity

$$\rho(P(x, \alpha_0), P_A(x|x_1, \ldots, x_i)) = \rho_{\alpha_0, A}(x_1, \ldots, x_l).$$

We shall define the closeness of densities in terms of the L^2 metric, i.e.,

$$\rho_{\alpha_0, A}(x_1, \ldots, x_l) = \left(\int (P(x, \alpha_0) - P_A(x|x_1, \ldots, x_l))^2 \, dx\right)^{1/2}. \quad (3.15)$$

Since the choice of the density $P_A(x|x_1, \ldots, x_l)$ depends on the sample $x_1, \ldots, x_l$, the quantity $\rho_{\alpha_0, A}(x_1, \ldots, x_l)$ is a random variable. We shall characterize the quality of the algorithm A by the mathematical expectation of $\rho^2_{\alpha_0, A}(x_1, \ldots, x_l)$:

$$R(\alpha_0, A) = \int \rho^2_{\alpha_0, A}(x_1, \ldots, x_l) P(x_1) \cdots P(x_l) \, dx_1 \cdots dx_l.$$

The smaller $R(\alpha_0, A)$ is, the better the algorithm is for estimating the density $P(x, \alpha_0)$ from a sample of size l.

Thus we have determined how the quality of an algorithm A designed for estimating a specific density $P(x, \alpha_0)$ should be measured. It is now necessary to agree on how to measure the quality of an algorithm earmarked for estimating an arbitrary density belonging to a given class $P(x, \alpha)$ (in our case the class of densities is defined up to values of a vector of parameters α). Two principles are used in statistical decision theory in such a situation: Bayes's principle and the minimax principle.

Bayes's principle asserts that the quality of an algorithm should be estimated as the mean quality over all the estimated densities. In order to estimate the mean value of an algorithm it is necessary to know how often any particular density belonging to $P(x, \alpha)$ will be estimated, i.e., in our case

† It will be shown in Section 6 that $\tilde{\tilde{p}} = (m + 1)/(l + 2)$ is a more precise estimator.

it is necessary to have information about the probability density $P(\alpha)$ of the vector of parameters α. In that case the quality of an algorithm is defined as

$$R_{\text{B}}(A) = \int R(\alpha, A)P(\alpha)\, d\alpha. \tag{3.16}$$

The smaller $R_{\text{B}}(A)$ is, the better the algorithm.

The minimax principle asserts that one must estimate the quality of an algorithm on the basis of the most unfavorable probability density $P(x, \alpha^*)$ for this algorithm. Here the densities which may be encountered in practice are not taken into account. It may therefore turn out that the quality of the algorithm is determined by a case which will never occur. The quality of an algorithm according to the minimax principle is defined as

$$R_{\text{mnx}}(A) = \sup_{\alpha} R(\alpha, A). \tag{3.17}$$

The smaller the value of $R_{\text{mnx}}(A)$, the better the algorithm.

§5 The Bayesian Algorithm for Density Estimation

We shall determine the structure of algorithms which assure the solution of the Bayesian estimation of density, i.e., which minimize the functional

$$R_{\text{B}}(A) = \int R(\alpha, A)P(\alpha)\, d\alpha.$$

From a sample $x_1, \ldots, x_l$, let a density which belongs to the class $P(x, \alpha)$ be estimated and the prior probability density $P(\alpha)$ be given. Utilizing Bayes's formula, we obtain

$$P(\alpha | x_1, \ldots, x_l) = \frac{P(x_1, \ldots, x_l | \alpha)P(\alpha)}{P(x_1, \ldots, x_l)},$$

which is the density of posterior probabilities $P(\alpha | x_1, \ldots, x_l)$ which characterizes the possibilities of realizations of various values of parameters α after the information about the sample $x_1, \ldots, x_l$ has been added to the prior information $P(\alpha)$. Here $P(x_1, \ldots, x_l | \alpha)$ is the conditional and $P(x_1, \ldots, x_l)$ is the unconditional density of occurrence of the sample $x_1, \ldots, x_l$:

$$P(x_1, \ldots, x_l) = \int P(x_1, \ldots, x_l | \alpha)P(\alpha)\, d\alpha.$$

Below we shall show that the posterior mean, i.e., the function

$$P_{\text{B}}(x | x_1, \ldots, x_l) = \int P(x, \alpha)P(\alpha | x_1, \ldots, x_l)\, d\alpha \tag{3.18}$$

is the solution of the Bayesian problem.

In general the density $P_B(x|x_1, \ldots, x_l)$ obtained as a result of averaging functions $P(x, \alpha)$ with respect to the measure $P(\alpha|x_1, \ldots, x_l)$ need not belong to the parametric family $P(x, \alpha)$ under consideration. Therefore, strictly speaking, the method for constructing the posterior mean (3.18) cannot actually be called the estimation of a function belonging to the class $P(x, \alpha)$.

Thus we obtain a function $\pi(x; x_1, \ldots, x_l)$ which minimizes the functional

$$R_B(\pi) = \int (P(x|\alpha) - \pi(x; x_1, \ldots, x_l))^2 \times P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha\, dx\, dx_1 \cdots dx_l. \tag{3.19}$$

Denote

$$r(x; x_1, \ldots, x_l) = \int (P(x|\alpha) - \pi(x; x_1, \ldots, x_l))^2 P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha.$$

Interchanging the order of integration in (3.19), we arrive at

$$R_B(\pi) = \int r(x; x_1, \ldots, x_l)\, dx\, dx_1 \cdots dx_l. \tag{3.20}$$

We now transform the function $r(x; x_1, \ldots, x_l)$:

$$\begin{aligned} r(x; x_1, \ldots, x_l) = &\int P^2(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha \\ &- 2\pi(x; x_1, \ldots, x_l) \int P(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha \\ &+ \pi^2(x; x_1, \ldots, x_l) \int P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha. \end{aligned} \tag{3.21}$$

Denote

$$\hat{P}(x|x_1, \ldots, x_l) = \frac{\int P(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha}{P(x_1, \ldots, x_l)},$$

where

$$P(x_1, \ldots, x_l) = \int P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha,$$

and rewrite the equality (3.21) in the form

$$\begin{aligned} r(x; x_1, \ldots, x_l) = &\int P^2(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha \\ &- \hat{P}^2(x|x_1, \ldots, x_l)P(x_1, \ldots, x_l) \\ &+ [\hat{P}(x|x_1, \ldots, x_l) - \pi(x; x_1, \ldots, x_l)]^2 P(x_1, \ldots, x_l). \end{aligned}$$

Substitute the expression for $r(x; x_1, \ldots, x_l)$ into (3.20). This results in a functional which can be expressed as the sum of two summands

$$R_B(\pi) = R_1 + R_2(\pi),$$

where

$$R_1 = \int [P^2(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha - P(x_1, \ldots, x_l)\hat{P}^2(x|x_1, \ldots, x_l)]\, dx\, dx_1 \cdots dx_l,$$

$$R_2(\pi) = \int [\hat{P}(x|x_1, \ldots, x_l) - \pi(x; x_1, \ldots, x_l)]^2\, dx\, dx_1 \cdots dx_l.$$

The first summand does not depend on $\pi(x; x_1, \ldots, x_l)$. Therefore minimization of $R_B(\pi)$ is equivalent to the minimization of the second summand $R_2(\pi)$. The minimum of this summand is zero and is attained if

$$\pi(x; x_1, \ldots, x_l) = \hat{P}(x|x_1, \ldots, x_l) \equiv P_B(x|x_1, \ldots, x_l).$$

In succeeding sections, for prior distributions $P(\alpha)$ Bayesian approximations of densities will be obtained. The construction of a Bayesian approximation for a fixed prior distribution $P(\alpha)$ depends on whether the expression (3.18) can be integrated analytically.

§6 Bayesian Estimators of Discrete Probability Distributions

In Section 3 the probability distribution function of the discrete independent features (3.12) and (3.13) was introduced. Here we shall show that, under minimal prior information concerning the values of the parameters $p^i(j)$, namely: for each i the parameters $p^i(1), \ldots, p^i(\tau_i)$ are uniformly distributed on the simplex

$$C_i = \left\{ p : \sum_{j=1}^{\tau_i} p^i(j) = 1, p^i(j) \geq 0 \right\}.$$

The Bayesian estimator of the probability distribution of discrete independent features equals

$$P_B(x) = \prod_{i=1}^{n} P_B(x^i),$$

where

$$P_B(x^i) = \begin{cases} \hat{p}^i(1) = \dfrac{m_i(1) + 1}{l + \tau_i}, \\ \vdots \\ \hat{p}^i(\tau_i) = \dfrac{m_i(\tau_i) + 1}{l + \tau_i}. \end{cases}$$

$m_i(j)$ is the number of vectors in the sample such that the rth coordinate takes the jth value, τ_i is the number of values taken by the ith coordinate, and l is the sample size.

We now obtain the Bayesian estimator of the probability distribution of the discrete independent features. For this purpose we compute the function

$$P_{\mathrm{B}}(x^i) = \frac{\int P(x^i|p)P(x_1^i, \ldots, x_l^i|p)P(p)\,dp}{\int P(x_1^i, \ldots, x_l^i|p)P(p)\,dp}. \tag{3.22}$$

In our case

$$P(x^i|p) = \begin{cases} p^i(1) & \text{for } x^i = c^i(1), \\ \vdots & \\ p^{(i)}(\tau_i) & \text{for } x^i = c^i(\tau_i). \end{cases}$$

First compute the denominator of (3.22). Since the sample is random and independent, we obtain

$$\int P(x_1^i, \ldots, x_l^i|p)P(p)\,dp = \frac{1}{v}\int_{C_i} \prod_{j=1}^{\tau_i} [p^i(j)]^{m_i(j)}\,dp^i(1)\cdots dp^i(\tau_i), \tag{3.23}$$

where v is the volume of the simplex C_i. It is known (see, e.g. [52]) that the definite integral (3.23) may be computed analytically:

$$P(x_1^i, \ldots, x_l^i) = \frac{1}{v}\frac{\Gamma(m_i(1)+1)\cdots\Gamma(m_i(\tau_i)+1)}{\Gamma(m_i(1)+\cdots+m_i(\tau_i)+\tau_i)}, \tag{3.24}$$

where $\Gamma(n)$ is the gamma function. For integer n this function is given by

$$\Gamma(n) = (n-1)!.$$

We now derive the numerator of the expression (3.22) for the case $x^i = c^i(k)$:

$$\begin{aligned} I_k^i &= \int_{C_i} P(x^i = c^i(k)|p)P(x_1^i, \ldots, x_l^i|p)P(p)\,dp \\ &= \frac{1}{v}\int_{C_i} p^i(k)\prod_{j=1}^{\tau_i}[p^i(j)]^{m_i(j)}\,dp^i(1)\cdots dp^i(\tau_i). \end{aligned}$$

The definite integral I_k^i is equal to (cf. [52])

$$I_k^i = \frac{1}{v}\frac{\Gamma(m_i(1)+1)\cdots\Gamma(m_i(\tau_i)+1)\Gamma(m_i(k)+2)}{\Gamma(m_i(1)+\cdots+m_i(\tau_i)+\tau_i+1)\Gamma(m_i(k)+1)}. \tag{3.25}$$

Dividing (3.25) by (3.24), we obtain

$$P_{\mathrm{B}}(x^i = c^i(k)) = \frac{\Gamma(m_i(k)+2)\Gamma(l+\tau_i)}{\Gamma(m_i(k)+1)\Gamma(l+\tau_i+1)} = \frac{m_i(k)+1}{l+\tau_i}.$$

Thus

$$P_{\mathrm{B}}(x^i) = \begin{cases} \hat{p}^i(1) = \dfrac{m_i(1)+1}{l+\tau_i} & \text{for } x^i = c^i(1), \\ \vdots & \\ \hat{p}^i(\tau_i) = \dfrac{m_i(\tau_i)+1}{l+\tau_i} & \text{for } x^i = c^i(\tau_i). \end{cases}$$

Since the features are independent, we have $P_{\mathrm{B}}(x) = \prod_{i=1}^{n} P_{\mathrm{B}}(x^i)$.

§7 Bayesian Estimators for the Gaussian (Normal) Density

We shall now obtain Bayesian estimators for the Gaussian (normal) density in some special cases of the prior distribution on the parameters. First we shall obtain Beyesian estimator for the univariate normal distribution $N(\mu, \sigma^2)$ under the assumption that the parameters μ and σ of this distribution are distributed uniformly on the rectangular region $0 \leq \sigma \leq \Pi$, $-T \leq \mu \leq T$. It turns out that if Π and T are sufficiently large, then the Bayesian estimators are equal to

$$P_{\text{B}}(x) = \frac{E(l)}{\sigma_{\text{emp}}}\left[1 + \frac{(x - x_{\text{emp}})^2}{(l+1)\sigma_{\text{emp}}^2}\right]^{-(l-1)/2}, \tag{3.26}$$

where

$$E(l) = \frac{\Gamma\left(\frac{l-1}{2}\right)}{\sqrt{(l+1)\pi}\,\Gamma\left(\frac{l}{2} - 1\right)},$$

$$x_{\text{emp}} = \frac{1}{l}\sum_{i=1}^{l} x_i, \qquad \sigma_{\text{emp}}^2 = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{emp}})^2.$$

(See the derivation below.)

Next we shall obtain the Bayesian estimators for the n-dimensional normal distribution for a special prior distribution on parameters μ and Δ (μ is an n-dimensional vector of the means and Δ is an $n \times n$ covariance matrix). It turns out that in this case the Bayesian approximation equals

$$P_{\text{B}}(x) = \frac{\bar{E}(l)}{|S|^{1/2}}\left[1 + \frac{(x - x_{\text{emp}})^{\text{T}} S^{-1}(x - x_{\text{emp}})}{l+1}\right]^{-(l+n)/2}, \tag{3.27}$$

where

$$\bar{E}(l) = \frac{\Gamma\left(\frac{l+n}{2}\right)}{((l+1)\pi)^{n/2}\Gamma(l/2)},$$

the vector x_{emp} is an estimator for the vector of the means:

$$x_{\text{emp}} = \frac{1}{l}\sum_{i=1}^{l} x_i,$$

and S is the empirical covariance matrix:

$$S = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{emp}})(x_i - x_{\text{emp}})^{\text{T}}.$$

Note that neither of the estimators (3.26) and (3.27) belongs to the normal class. However, it is easy to verify that in both cases

$$P_{\mathrm{B}}(x) \underset{l\to\infty}{\longrightarrow} N(\mu, \Delta).$$

as $l \to \infty$.

Yet another remark: In order to calculate explicitly the Bayesian estimators of a multidimensional normal distribution (see Section 7.2 below) it was necessary to consider a special prior distribution on the parameters which differs from the uniform one (used in the univariate case; see Section 7.1 below). However, the Bayesian estimators for the univariate case obtained from (3.27) by setting $n = 1$ is close to the one obtained assuming the uniform distribution on the parameters in the univariate case (3.26).

7.1 Bayesian Estimator for the Univariate Normal Distribution

Let the variable x be distributed according to the normal distribution

$$P(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Moreover, let the prior distribution of parameters μ and σ be uniform in the rectangle $0 \le \sigma \le \Pi$, $-T \le \mu \le T$; since the sample $x_1, \ldots, x_l$ is random and independent, we have

$$P(x_1, \ldots, x_l; \mu, \sigma) = \frac{1}{(2\pi)^{l/2}\sigma^l} \exp\left\{-\frac{\sum_{i=1}^{l}(x_i-\mu)^2}{2\sigma^2}\right\}.$$

In view of (3.18) the Bayesian estimator of the probability density is equal to

$$P_{\mathrm{B}}(x) = \left(\frac{1}{2T\Pi}\frac{1}{(2\pi)^{(l+1)/2}}\int_{-T}^{T}\int_{0}^{\Pi}\frac{1}{\sigma^{l+1}}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{l}(x_i-\mu)^2+(x-\mu)^2\right)\right\}d\mu\, d\sigma\right)$$

$$\times\left(\frac{1}{2T\Pi}\frac{1}{(2\pi)^{l/2}}\int_{-T}^{T}\int_{0}^{\Pi}\frac{1}{\sigma^{l}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{l}(x_i-\mu)^2\right\}d\mu\, d\sigma\right)^{-1}. \tag{3.28}$$

We shall assume that the intervals $[-T, T]$ and $[0, \Pi]$ are so large that the limits of integration in (3.28) may be extended to $(-\infty, \infty)$ and $(0, \infty)$ respectively. This can evidently be done if $l \ge 2$. In this case the integrals in (3.28) are convergent. We compute the numerator of (3.28):

$$I(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{1}{\sigma^{l+1}}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{l}(x_i-\mu)^2+(x-\mu)^2\right)\right\}d\mu\, d\sigma. \tag{3.29}$$

Denote

$$T(\mu) = \sum_{i=1}^{l}(x_i-\mu)^2 + (x-\mu)^2, \qquad y = \frac{\sqrt{T(\mu)}}{\sigma}.$$

Then the integral (3.29) becomes

$$I(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{0}^{\infty} \frac{y^{l-1}}{T^{l/2}(\mu)} \exp\{-\tfrac{1}{2}y^2\}\, dy\, d\mu$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{d\mu}{T^{l/2}(\mu)} \int_{0}^{\infty} y^{l-1} \exp\left\{-\frac{y^2}{2}\right\} dy.$$

Denoting

$$c(l) = \int_{0}^{\infty} y^{l-1} \exp\left\{-\frac{y^2}{2}\right\} dy,$$

where $c(l)$ depends on neither μ nor on σ, this integral can be rewritten as

$$I(x) = \frac{c(l)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{d\mu}{T^{l/2}(\mu)}.$$

We now transform the expression for $T(\mu)$. For this purpose we note that

$$\sum_{i=1}^{l} (x_i - \mu)^2 = l\sigma_{\text{emp}}^2 + l(\mu - x_{\text{emp}})^2,$$

where

$$x_{\text{emp}} = \frac{1}{l} \sum_{i=1}^{l} x_i, \qquad \sigma_{\text{emp}}^2 = \frac{1}{l} \sum_{i=1}^{l} (x_i - x_{\text{emp}})^2.$$

The expression for $T(\mu)$ is transformed analogously to yield

$$T(\mu) = l\sigma_{\text{emp}}^2 + l(\mu - x_{\text{emp}})^2 + (x - \mu)^2.$$

Now set

$$\bar{x} = \frac{x_{\text{emp}} l + x}{l + 1}$$

and rewrite $T(\mu)$:

$$T(\mu) = l\sigma_{\text{emp}}^2 + \frac{l}{l+1}(x - x_{\text{emp}})^2 + (\bar{x} - \mu)^2(l + 1).$$

We now write the integral $I(x)$ in the form

$$I(x) = \frac{c(l)}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{d\mu}{\left[l\sigma_{\text{emp}}^2 + \frac{l}{l+1}(x - x_{\text{emp}})^2 + (\bar{x} - \mu)^2(l + 1)\right]^{l/2}}$$

$$= \frac{c(l)}{\sqrt{2\pi(l+1)}} \left(l\sigma_{\text{emp}}^2 + \frac{l(x - x_{\text{emp}})^2}{(l+1)}\right)^{-(l-1)/2} \int_{-\infty}^{\infty} \frac{dz}{(1 + z^2)^{l/2}}.$$

Observe now that the integrand is independent of the parameters. We thus have

$$I(x) = c'(l, \sigma_{\text{emp}}) \left(1 + \frac{(x - x_{\text{emp}})^2}{(l+1)\sigma_{\text{emp}}^2}\right)^{-(l-1)/2} \tag{3.30}$$

To obtain a Bayesian estimator it is required only to normalize the expression (3.30):

$$P_{\mathrm{B}}(x) = \frac{I(x)}{\int_{-\infty}^{\infty} I(x)\,dx}.$$

It is known (cf. [52]) that the integral in the denominator equals the following expression:

$$\int_{-\infty}^{+\infty} I(x)\,dx = c''(l, \sigma_{\mathrm{emp}}) \int_{-\infty}^{+\infty} \frac{dx}{\left(1 + \dfrac{(x - x_{\mathrm{emp}})^2}{(l+1)\sigma_{\mathrm{emp}}^2}\right)^{(l-1)/2}}$$

$$= \frac{c''(l, \sigma_{\mathrm{emp}})\,\sigma_{\mathrm{emp}}\sqrt{l+1}\cdot\Gamma\left(\dfrac{1}{2}\right)\Gamma\left(\dfrac{l}{2} - 1\right)}{\Gamma\left(\dfrac{l-1}{2}\right)}.$$

Denote

$$E(l) = \frac{\Gamma\left(\dfrac{l-1}{2}\right)}{\sqrt{l+1}\,\Gamma\left(\dfrac{1}{2}\right)\Gamma\left(\dfrac{l}{2} - 1\right)} = \frac{\Gamma\left(\dfrac{l-1}{2}\right)}{\sqrt{\pi(l+1)}\,\Gamma\left(\dfrac{l}{2} - 1\right)}.$$

Thus

$$P_{\mathrm{B}}(x) = \frac{E(l)}{\sigma_{\mathrm{emp}}}\left(1 + \frac{(x - x_{\mathrm{emp}})^2}{(l+1)\sigma_{\mathrm{emp}}^2}\right)^{-(l-1)/2}.$$

7.2 Bayesian Estimator for the n-dimensional Normal Distribution

To obtain the Bayesian estimator for the n-dimensional normal distribution, the following two facts from the theory of multidimensional normal distributions are used:

(1) The convolution of two multidimensional normal distributions $N(0, \Delta)$ and $N(\mu, \gamma\Delta)$, where γ is a positive number, is the normal distribution $N(\mu, (1+\gamma)\Delta)$. In other words the equality

$$\int_{E_n} N(\mu - t, \gamma\Delta) \cdot N(t, \Delta)\,dt = N(\mu, (1+\gamma)\Delta)$$

is valid (see [4]).

(2) The distribution of empirical estimators S of the covariance matrix Δ given by the formula

$$S = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\mathrm{emp}})(x_i - x_{\mathrm{emp}})^{\mathrm{T}}, \qquad x_{\mathrm{emp}} = \frac{1}{l}\sum_{i=1}^{l} x_i,$$

is expressed by the Wishart distribution (see [5]):

$$W_{l,n}(S; \Delta) = \begin{cases} C_{n,l}|\Delta|^{-(l-1)/2}|S|^{(l-n-2)/2}\exp\left\{-\dfrac{l}{2}\mathrm{Sp}[\Delta^{-1}S]\right\} & \text{for } |S| > 0, \\ 0 & \text{for } |S| \le 0, \end{cases}$$

where it is assumed that $l > n + 1$, $\mathrm{Sp}\|a_{ij}\| = \sum_{i=1}^{n} a_{ii}$. The quantity $C_{n,l}$ is a constant and equals

$$C_{n,l} = \left(\left(\frac{l}{2}\right)^{-(l-1)n/2} \pi^{n(n-1)/4} \prod_{i=1}^{n} \Gamma\left(\frac{l-i}{2}\right)\right)^{-1}. \tag{3.31}$$

Since the Wishart distribution sums to 1, we have

$$\int_{|S|>0} |S|^{(l-n-2)/2} \exp\left\{-\frac{l}{2}\mathrm{Sp}[\Delta^{-1}S]\right\} dS = \frac{1}{C_{n,l}} |\Delta|^{(l-1)/2}. \tag{3.32}$$

We now derive the Bayesian estimator. Denote the matrix Δ^{-1} by $\mathscr{D}$. Clearly $|\Delta| = 1/|\mathscr{D}|$. Let the prior distribution of parameters μ and Δ of an n-dimensional normal distribution $N(\mu, \Delta)$ be defined in the form

$$P_{a,A}(\mu, \mathscr{D}) = P_a(\mu|\mathscr{D}) \cdot P_A(\mathscr{D}),$$

where the vector μ is distributed according to the normal distribution

$$P_a(\mu|\mathscr{D}) = c_1 |\mathscr{D}|^{1/2} \exp\left\{-\frac{\omega}{2}(\mu - a)^{\mathrm{T}}\mathscr{D}(\mu - a)\right\};$$

here c_1 is a constant, $\omega > 0$ is a number, a is a vector, and $\mathscr{D}$ is a matrix distributed according to the Wishart distribution:

$$P_A(\mathscr{D}) = \begin{cases} C_{n,v}|\omega A|^{(v-1)/2}|\mathscr{D}|^{(v-n-2)/2} \exp\left\{-\dfrac{v\omega}{2}\mathrm{Sp}[A\mathscr{D}]\right\} & \text{for } |\mathscr{D}| > 0, \\ 0 & \text{for } |\mathscr{D}| \le 0. \end{cases}$$

Here $v > n + 2$ is a constant, A is a matrix. Observe that

$$\mathrm{Sp}[\mathscr{D}xx^{\mathrm{T}}] = \mathrm{Sp}[xx^{\mathrm{T}}\mathscr{D}] = x^{\mathrm{T}}\mathscr{D}x, \tag{3.33}$$

where $\mathscr{D}$ is a symmetric matrix and x is a column vector. We now write the joint density $P(x_1, \ldots, x_l|\mu, \mathscr{D})$ for a random independent sample $x_1, \ldots, x_l$:

$$\begin{aligned} P(x_1, \ldots, x_l|\mu, \mathscr{D}) &= c_2|\mathscr{D}|^{l/2} \exp\left\{\frac{-\sum_{i=1}^{l}(x_i - \mu)^{\mathrm{T}}\mathscr{D}(x_i - \mu)}{2}\right\} \\ &= c_2|\mathscr{D}|^{l/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[l\mathscr{D}S + l\mathscr{D}(x_{\mathrm{emp}} - \mu)(x_{\mathrm{emp}} - \mu)^{\mathrm{T}}]\}. \end{aligned}$$

Here and below c_0, c_1, c_2, and c_3 are constants which are determined by normalizing conditions. In view of Bayes's formula the posterior density $P(\mu, \mathscr{D}|x_1, \ldots, x_l)$ equals

$$P(\mu, \mathscr{D}|x_1, \ldots, x_l) = c_0 P(x_1, \ldots, x_l|\mu, \mathscr{D}) P_a(\mu|\mathscr{D}) P_A(\mathscr{D}). \tag{3.34}$$

Compute the right-hand side of (3.34):

$$\begin{aligned} &P(\mu, \mathscr{D}|x_1, \ldots, x_l) \\ &\quad = c_0|\mathscr{D}|^{l/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[l\mathscr{D}S + l\mathscr{D}(x_{\mathrm{emp}} - \mu)(x_{\mathrm{emp}} - \mu)^{\mathrm{T}}]\} \\ &\qquad \times c_1|\mathscr{D}|^{l/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[\mathscr{D}\omega(\mu - a)(\mu - a)^{\mathrm{T}}]\} \\ &\qquad \times c_2 \cdot C_{n,v}|\mathscr{D}|^{(v-n-2)/2}|\omega A|^{(v-1)/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[v\mathscr{D}A\omega]\} \\ &\quad = c_3|\mathscr{D}|^{(l+v-n-1)/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[l\mathscr{D}S + l\mathscr{D}(x_{\mathrm{emp}} - \mu)(x_{\mathrm{emp}} - \mu)^{\mathrm{T}} \\ &\qquad\qquad + \omega\mathscr{D}(\mu - a)(\mu - a)^{\mathrm{T}} + v\omega\mathscr{D}A]\}. \end{aligned} \tag{3.35}$$

Transforming the expression in the exponent of (3.35), we obtain

$$\mathscr{D}(lS + l(x_{\text{emp}} - \mu)(x_{\text{emp}} - \mu)^{\mathrm{T}} + \omega(\mu - a)(\mu - a)^{\mathrm{T}} + v\omega A) = \mathscr{D}[(l + \omega)(\mu - b)(\mu - b)^{\mathrm{T}} + (l + v)B],$$

where the notation

$$b = \frac{lx_{\text{emp}} + a\omega}{l + \omega}, \qquad B = \frac{\left(lS + \omega v A + \dfrac{l\omega}{l + \omega}(x_{\text{emp}} - a)(x_{\text{emp}} - a)^{\mathrm{T}}\right)}{l + v} \tag{3.36}$$

is used. Using this notation we rewrite (3.35):

$$P(\mu, \mathscr{D} | x_1, \ldots, x_l) = c_3 |\mathscr{D}|^{(l+v-n-1)/2} \times \exp\{-\tfrac{1}{2}\operatorname{Sp}[\mathscr{D}((l + \omega)(\mu - b)(\mu - b)^{\mathrm{T}} + (l + v)B)]\}. \tag{3.37}$$

The normalizing condition allows us to determine the constant c_3:

$$\begin{aligned} c_3^{-1} &= \int |\mathscr{D}|^{(l+v-n-2)/2} \exp\left\{-\frac{l + v}{2}\operatorname{Sp}[\mathscr{D}B]\right\} d\mathscr{D} \\ &\quad \times \int |\mathscr{D}|^{1/2} \exp\left\{-\frac{l + \omega}{2}\operatorname{Sp}[\mathscr{D}(\mu - b)(\mu - b)^{\mathrm{T}}]\right\} d\mu \\ &= \left(\frac{2\pi}{l + \omega}\right)^{n/2} (C_{n,l+v} |(l + v)B|^{(l+v-1)/2})^{-1}. \end{aligned}$$

The outer integral was computed utilizing equality (3.32). Finally we obtain the Bayesian estimator

$$\begin{aligned} P_{\mathrm{B}}(x) &= \int P(x | \mu, \mathscr{D}) P(\mu, \mathscr{D} | x_1, \ldots, x_l)\, d\mu\, d\mathscr{D} \\ &= \int (2\pi)^{-n/2} |\mathscr{D}|^{1/2} \exp\left\{-\frac{1}{2}(x - \mu)^{\mathrm{T}}\mathscr{D}(x - \mu)\right\} c_3 |\mathscr{D}|^{(l+v-n-1)/2} \\ &\quad \times \exp\left\{-\frac{l + \omega}{2}(\mu - b)^{\mathrm{T}}\mathscr{D}(\mu - b)\right\} \exp\left\{-\frac{l + v}{2}\operatorname{Sp}[\mathscr{D}B]\right\} d\mu\, d\mathscr{D} \\ &= \left(\frac{2\pi}{l + \omega}\right)^{n/2} \int c_3 |\mathscr{D}|^{(l+v-n-2)/2} \exp\left\{-\frac{l + v}{2}\operatorname{Sp}[\mathscr{D}B]\right\} d\mathscr{D} \\ &\quad \times \int (2\pi)^{-n/2} |\mathscr{D}|^{1/2} (2\pi)^{-n/2} |(l + \omega)\mathscr{D}|^{1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}(x - \mu)^{\mathrm{T}}\mathscr{D}(x - \mu)\right\} \exp\left\{-\frac{l + \omega}{2}(\mu - b)^{\mathrm{T}}\mathscr{D}(\mu - b)\right\} d\mu. \end{aligned}$$

Observe that the inner integral with respect to μ is a convolution of two normal distributions; we thus obtain

$$P_{\mathrm{B}}(x) = c_3 \int (l + \omega + 1)^{-n/2} |\mathscr{D}|^{(l+v-n-1)/2} \times \exp\left\{-\tfrac{1}{2}\operatorname{Sp}\left[\mathscr{D}\left(B(l + v) + \frac{l + \omega}{l + \omega + 1}(x - b)(x - b)^{\mathrm{T}}\right)\right]\right\} d\mathscr{D}. \tag{3.38}$$

In view of (3.32) we have

$$P_B(x) = \frac{c_3(l+\omega+1)^{-n/2}}{C_{n,l+v+1}} \times \left|(l+v)B + \frac{l+\omega}{l+\omega+1}(x-b)(x-b)^T\right|^{-(l+v)/2}$$

$$= \left(\frac{l+\omega}{l+\omega+1}\frac{1}{2\pi}\right)^{n/2} \frac{C_{n,l+v}}{C_{n,l+v+1}} \times \frac{|(l+v)B|^{(l+v-1)/2}}{\left|(l+v)B + \frac{l+\omega}{l+\omega+1}(x-b)(x-b)^T\right|^{(l+v)/2}}. \tag{3.39}$$

We now transform the expression (3.39):

$$P_B(x) = \left(\frac{1}{\pi}\frac{l+\omega}{l+\omega+1}\right)^{n/2} \frac{\Gamma\left(\frac{l+v}{2}\right)}{\Gamma\left(\frac{l+v-n}{2}\right)} \times \frac{|(l+v)\cdot B|^{-1/2}}{\left|I + \frac{l+\omega}{l+\omega+1}\frac{1}{l+v}(x-b)(x-b)^T B^{-1}\right|^{(l+v)/2}}. \tag{3.40}$$

In the denominator of this expression I is the unit matrix. Observe that the matrix $(x-b)(x-b)^T$ and hence the matrix $(x-b)(x-b)^T B^{-1}$ are of rank 1. Thus only one of its eigenvalues is different from zero, which implies that the denominator of (3.40) is equal to

$$\left|I + \frac{l+\omega}{l+\omega+1}\frac{1}{l+v}(x-b)(x-b)^T B^{-1}\right|^{(l+v)/2} = \left(1 + \frac{l+\omega}{l+\omega+1}\frac{1}{l+v}(x-b)^T B^{-1}(x-b)\right)^{(l+v)/2}.$$

Thus we finally obtain

$$P_B(x) = \left(\frac{1}{\pi}\frac{l+\omega}{l+\omega+1}\right)^{n/2} \frac{\Gamma\left(\frac{l+v}{2}\right)}{\Gamma\left(\frac{l+v-n}{2}\right)} \times \frac{|(l+v)B|^{-1/2}}{\left(1 + \frac{l+\omega}{l+\omega+1}\frac{1}{l+v}(x-b)^T B^{-1}(x-b)\right)^{(l+v)/2}}.$$

We now assign specific values for v and ω in order that under the conditions of the scheme we shall obtain the most general (undetermined) prior conditions:

(1) $v = n + \varepsilon$ ($\varepsilon > 0$). This condition is necessary for integrating Wishart's distribution.

(2) $\omega \to 0$, $\varepsilon \to 0$. This condition assures that each of the elements of the matrix A tends to zero.

Then in view of (3.36) we obtain that $b \to x_{\text{emp}}$, $(l + v)B \to lS$, whence

$$P_{\text{B}}(x) = \left(\frac{1}{(l+1)\pi}\right)^{n/2} \frac{\Gamma\left(\frac{l+n}{2}\right)}{\Gamma\left(\frac{l}{2}\right)} \frac{|S|^{-1/2}}{\left(1 + \frac{1}{l+1}(x - x_{\text{emp}})^{\text{T}} S^{-1}(x - x_{\text{kmp}})\right)^{(l+n)/2}}.$$

Finally for the one-dimensional case (setting $n = 1$) we obtain

$$P_{\text{B}}(x) = \sqrt{\frac{1}{\pi(l+1)}} \frac{1}{\sigma_{\text{emp}}} \frac{\Gamma\left(\frac{l+1}{2}\right)}{\Gamma\left(\frac{l}{2}\right)} \frac{1}{\left(1 + \frac{1}{l+1}\frac{(x - x_{\text{emp}})^2}{\sigma_{\text{emp}}^2}\right)^{(l+1)/2}}.$$

§8 Unbiased Estimators

In the preceding sections, the Bayesian estimators of a probability density for special prior distributions on parameters were obtained. However, in practical problems the prior distribution is usually unknown. The minimax scheme of estimating the density may lead to overly imprecise results. It would therefore be desirable to find a sufficiently reliable method of estimating densities which is not connected with the Bayesian approach. How can this be done?

Assume that there exists a method of estimating densities which is best not only on the average (this corresponds to the Bayesian criterion), but also the best for estimating each specific density. For this uniformly best method to exist it must be independent of the prior distribution imposed on the density.

Unfortunately there is no such (uniformly best) method of estimation in the class of all possible methods. Indeed there exists a trivial algorithm which estimates the density to have the same fixed values of parameters independently of the sample. Such an algorithm estimates a single density with complete precision, while it is a poor estimator for all the other ones. This estimator is of course the best for its own density.

However, while there is no uniformly best method in the class of all possible estimation methods, there may perhaps exist such a method in a more restricted class. This prompts the idea of restricting the class of all possible methods of density estimation and attempting to find the best method within the class. It turns out that if we restrict the class of estimators to the so-called *unbiased estimators of density*, then the problem of finding a uniformly best one admits a solution.

Definition. We say that the function $\pi(x; x_1, \ldots, x_l)$ is an *unbiased estimator* of the density $P(x, \alpha^*)$ belonging to the class $P(x, \alpha)$ constructed from a sample $x_1, \ldots, x_l$ of size l obtained according to distribution $P(x, \alpha^*)$ if

the mathematical expectation of the estimator $\pi(x_1, \ldots, x_l)$ equals the density $P(x, \alpha^*)$, i.e., if for any $P(x, \alpha^*)$ belonging to $P(x, \alpha)$ the equality

$$M_{\alpha^*}\pi(x; x_1, \ldots, x_l) = P(x, \alpha^*)$$

is valid.

Note that the unbiasedness property has no value on its own and it is introduced solely to narrow down the class of possible estimators. The reason why the class of unbiased estimators is widely used in statistics is that this class is accessible to analysis.

What is the meaning of this accessibility? We write once again the definition of an unbiased estimator:

$$\int \pi(x; x_1, \ldots, x_l)P(x_1, \ldots, x_l; \alpha)\, dx_1 \cdots dx_l = P(x, \alpha). \tag{3.41}$$

This expression not only determines unbiased density estimators, but indicates a method for their construction: the set of unbiased estimators is the set of solutions of Fredholm's equation of type I. However, to obtain a solution of Equation (3.41) is usually a difficult problem. It was shown in Chapter 1 that even in the case when the solution of Fredholm's equation is unique, its numerical solution is an ill-posed problem. Therefore one can obtain unbiased estimators of the density $P(x, \alpha)$ only if Equation (3.41) can be solved analytically.

In Section 10 an optimal unbiased estimator of density for a multivariate normal distribution will be derived. Before proceeding to construct this estimator, we note that in Chapter 2 a more general problem of density estimation in the class of continuous functions was also reduced to a solution of Fredholm's equation of type I. In this case a special problem—obtaining an unbiased estimator of a density known up to its parameters—is reduced to Fredholm's equation.

The substantial difference between these two situations is that in the general case considered in Chapter 2 the right-hand side of Fredholm's equation of type I is known up to the error term. Here, however, it is given precisely.

§9 Sufficient Statistics

The construction of the optimal unbiased estimator is possible in terms of the so-called *sufficient statistics*. Up until now, when studying estimators we assumed that the estimator of a density is of the form $\pi(x; x_1, \ldots, x_l)$, i.e., the estimator is a function of $l + 1$ vectors: the vector x and l vector-valued variables $x_1, \ldots, x_l$. Fixing the last l variables we obtained a specific form of the estimated density.

However, such a method of expressing the density estimator is not quite convenient. Evidently $\pi(x; x_1, \ldots, x_l)$ should not depend on the order of

the vectors $x_1, \ldots, x_l$ of the sample. Moreover, for another sample size, say $l + 1$, it is necessary to give a new function (of dimensionality $l + 2$).

Therefore it would be desirable to find k characteristics of the sample

$$t_i = f_i(x_1, \ldots, x_l), \qquad i = 1, \ldots, k,$$

such that, first of all, the information concerning the density contained in the sample $x_1, \ldots, x_l$ would be included in these k numbers, and secondly, that the number of necessary characteristics k would depend not on the sample size but on the features of the class of estimated densities. It would be desirable to obtain an unbiased estimator $\pi^*(x; t_1, \ldots, t_k)$ in terms of these characteristics of the sample. Sufficient statistics indeed serve this purpose (see [58]).

Definition. We say that the functions $t_i = f_i(x_1, \ldots, x_l)$ are *sufficient statistics* for the density $P(x, \alpha)$ if the joint density $P(x_1, \ldots, x_l; \alpha)$ of the sample can be represented in the form

$$P(x_1, \ldots, x_l; \alpha) = P_1(t_1, \ldots, t_k; \alpha) P_2(x_1, \ldots, x_l).$$

In other words, the joint density $P(x_1, \ldots, x_l; \alpha)$ is decomposed into the product of two terms. One of them, $P_2(\cdot)$, does not depend on the parameter α, while the other involving α depends only on the values $t_1, \ldots, t_k$ (but not on the sample $x_1, \ldots, x_l$).

It is easy to verify that for an n-dimensional normal distribution the following $n(n + 3)/2$ quantities serve as sufficient statistics:

$$t = \frac{1}{l} \sum_{j=1}^{l} x_j, \quad t = (t_1, \ldots, t_n)^{\mathrm{T}} \qquad (n \text{ values});$$

$$\|t_{ij}\| = \sum_{r=1}^{l} (x_r - t)(x_r - t)^{\mathrm{T}} \qquad \left(\frac{n(n+1)}{2} \text{ values}\right).$$

Indeed, for an n-dimensional normal distribution we have

$$\begin{aligned}
&P(x_1, \ldots, x_l; \mu, \Delta) \\
&\quad = \frac{1}{(2\pi)^{nl/2} |\Delta|^{l/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{l} (x_i - \mu)^{\mathrm{T}} \Delta^{-1} (x_i - \mu) \right\} \\
&\quad = (2\pi)^{-nl/2} |\Delta|^{-l/2} \exp\left\{ -\tfrac{1}{2} \operatorname{Sp}\left[\Delta^{-1} \sum_{i=1}^{l} (x_i - \mu)(x_i - \mu)^{\mathrm{T}} \right\} \right] \\
&\quad = (2\pi)^{-nl/2} |\Delta|^{-l/2} \\
&\qquad \times \exp\left\{ -\tfrac{1}{2} \operatorname{Sp}\left[\Delta^{-1} \left(\sum_{i=1}^{l} (x_i - t)(x_i - t)^{\mathrm{T}} + l(t - \mu)(t - \mu)^{\mathrm{T}} \right) \right] \right\} \\
&\quad = (2\pi)^{-nl/2} |\Delta|^{-l/2} \exp\{ -\tfrac{1}{2} \operatorname{Sp}[\Delta^{-1}(\|t_{ij}\| + l(t - \mu)(t - \mu)^{\mathrm{T}})] \}.
\end{aligned}$$

In the derivation the equality $z^{\mathrm{T}} B z = \operatorname{Sp}[z z^{\mathrm{T}} B]$ was used.

Thus we seek an estimator of the density as a function of sufficient statistics.

The remarkable feature of unbiased estimators $\pi^*(x; t_1, \ldots, t_k)$ is that they are in some sense always at least as good as the estimators $\pi(x; x_1, \ldots, x_l)$.

Theorem (cf. [35, 58]). *For any estimator $\pi(x; x_1, \ldots, x_l)$ there exists an estimator $\pi^*(x; t_1, \ldots, t_k)$ such that for any density belonging to $P(x, \alpha)$ the mathematical expectations of the estimators are the same:*

$$M\pi^*(x; t_1, \ldots, t_k) = M\pi(x; x_1, \ldots, x_l) = \pi(x),$$

but the variance $\pi^*(x; t_1, \ldots, t_k)$ is not larger than the variance of the estimator $\pi^*(x; x_1, \ldots, x_l)$, i.e.,

$$M(\pi(x) - \pi^*(x; t_1, \ldots, t_k))^2 \leq M(\pi(x) - \pi(x; x_1, \ldots, x_l))^2.$$

It follows from this theorem that the class of unbiased estimators—expressed in terms of a sufficient statistic—contains the best one.

§10 Computing the Best Unbiased Estimator

We shall construct the best unbiased estimator of the density for a multidimensional normal distribution. Here we utilize the fact that for distributions of the exponential type there exists a unique unbiased estimator expressed in terms of sufficient statistics [26, 35]. In other words there exists a unique solution for Fredholm's equation of type I,

$$\int \pi^*(x; t_1, \ldots, t_k) P(t_1, \ldots, t_k; \alpha)\, dt_1, \ldots, dt_k = P(x, \alpha), \tag{3.42}$$

where $P(x, \alpha)$ is the normal distribution and $P(t_1, \ldots, t_k; \alpha)$ is the probability density of its sufficient statistics.

According to the theorem cited in the preceding section, the solution of Equation (3.42), in view of its uniqueness, is the best unbiased estimator of the density of a multidimensional normal distribution.

We shall show that an unbiased estimator of an n-dimensional normal density is

$$P_{\text{unb}}(x) = \frac{\Gamma\left(\frac{l-1}{2}\right)}{[(l-1)\pi]^{n/2}\Gamma\left(\frac{l-n-1}{2}\right)|S|^{1/2}} \times \left[1 - \frac{(x - x_{\text{emp}})^{\mathrm{T}} S^{-1}(x - x_{\text{emp}})}{l-1}\right]_+^{(l-n-3)/2}.$$

Here $x_{\text{emp}} = (1/l)\sum_{i=1}^{l} x_i$ is the vector of the means,

$$S = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{emp}})(x_i - x_{\text{emp}})^{\mathrm{T}}$$

is the empirical estimator of the covariance matrix Δ, and $[z]_+$ denotes

$$[z]_+ = \begin{cases} z & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

In deriving the best unbiased estimator of an n-dimensional density we shall utilize Bayes's formula

$$\varphi(x_l|t) = \frac{q(x_l, t; \alpha)}{P(t; \alpha)}, \tag{3.43}$$

where $t = (t_1, \ldots, t_k)^{\mathrm{T}}$, $x_l = (x_l^1, \ldots, x_l^n)^{\mathrm{T}}$, the density $q(x_l, t; \alpha)$ defines the distribution of statistics x_l and t, $P(t, \alpha)$ is the distribution of t, and $\varphi(x_l|t)$ is the conditional density. We shall show that the conditional density (3.43) is an unbiased estimator of the density $P(x, \alpha)$. Indeed,

$$\int \varphi(x_l|t)P(t; \alpha)\, dt = \int q(x, t; \alpha)\, dt = P(x, \alpha).$$

And since the unbiased estimator expressed in terms of sufficient statistics is unique, $\varphi(x|t)$ is the best unbiased estimator.

We now compute $\varphi(x|t)$. First we shall find $q(x, t; \alpha)$. For a normal distribution of the occurrence of vector x we have

$$q(x, t; \alpha) = q(x, x_{\text{emp}}, S; \mu, \Delta),$$

where

$$x = x_l, \qquad x_{\text{emp}} = \frac{1}{l}\sum_{i=1}^{l} x_i, \qquad S = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{emp}})(x_i - x_{\text{emp}})^{\mathrm{T}}.$$

Let the vectors $x_1, \ldots, x_l$ which form the triples x, x_{emp}, S appear randomly and independently according to the density $N(\mu, \Delta)$.

Consider vectors $y_1, \ldots, y_l$ obtained from $x_1 - \mu, \ldots, x_l - \mu$ by an orthogonal transformation

$$\mathscr{L} = \begin{bmatrix} c_{11} & \cdots & c_{1\,l-1} & 0 \\ \vdots & & \vdots & \\ c_{l-2\,1} & \cdots & c_{l-2\,l-1} & 0 \\ \dfrac{1}{\sqrt{l-1}} & \cdots & \dfrac{1}{\sqrt{l-1}} & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Vectors $y_1, \ldots, y_l$ are distributed independently according to the $N(0, \Delta)$ distribution. The following relations are valid:

$$x_l = y_l + \mu, \qquad x_{\text{emp}} = \frac{\sqrt{l-1}}{l}\, y_{l-1} + \frac{y_l}{l} + \mu.$$

We now express the matrix S in terms of the vectors $y_1, \ldots, y_l$. For this purpose we utilize the representation

$$
\begin{aligned}
S = {} & \frac{1}{l}\sum_{i=1}^{l-1}(x_i-\mu)(x_i-\mu)^{\mathrm{T}} \\
& + \frac{(x_l-\mu)(x_l-\mu)^{\mathrm{T}}}{l} - \frac{l-1}{l}\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right]\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right]^{\mathrm{T}} \\
& - \frac{l-1}{l}\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right](x_l-\mu)^{\mathrm{T}} - \frac{l-1}{l}(x_l-\mu)\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right]^{\mathrm{T}} \\
& - \frac{1}{l^2}(x_l-\mu)(x_l-\mu)^{\mathrm{T}},
\end{aligned}
$$

and the fact that the transformation $\mathscr{L}$ satisfies

$$
\sum_{i=1}^{l-1}(x_i-\mu)(x_i-\mu)^{\mathrm{T}} = \sum_{i=1}^{l-1} y_i y_i^{\mathrm{T}}.
$$

We thus obtain

$$
S = \frac{1}{l}\sum_{i=1}^{l-2} y_i y_i^{\mathrm{T}} + \left(\frac{y_{l-1}-\sqrt{l-1}\,y_l}{l}\right)\cdot\left(\frac{y_{l-1}-\sqrt{l-1}\,y_l}{l}\right)^{\mathrm{T}}.
$$

Denote

$$
\mathscr{D} = \frac{1}{l}\sum_{i=1}^{l-2} y_i y_i^{\mathrm{T}}.
$$

Observe that vectors $y_1, \ldots, y_l$ are distributed according to the normal distribution $N(0, \Delta)$. Moreover the variables y_{l-1}, y_l, and $\mathscr{D}$ are independent. Since y_{l-1}, y_l are distributed according to the normal distribution and $\mathscr{D}$ has a Wishart distribution, the joint distribution $P(y_{l-1}, y_l, \mathscr{D}; 0, \Delta)$ equals

$$
P(y_{l-1}, y_l, \mathscr{D}; 0, \Delta) = P(y_{l-1}; 0, \Delta)\, P(y_l; 0, \Delta)\, W_{l-1}(\mathscr{D}; \Delta), \tag{3.44}
$$

where $W_{l-1}(\mathscr{D}, \Delta)$ is the Wishart distribution:

$$
W_{l-1}(\mathscr{D}, \Delta) = \begin{cases} C_{n,l-1}\dfrac{|\mathscr{D}|^{(l-n-3)/2}\exp\{-\frac{1}{2}\operatorname{Sp}[\Delta^{-1}\mathscr{D}]\}}{|\Delta|^{(l-2)/2}} & \text{for } |\mathscr{D}| > 0, \\ 0 & \text{for } |\mathscr{D}| \le 0, \end{cases}
$$

and $C_{n,l}$ is a constant defined in (3.31).

We now express $P(y_{l-1}, y_l, \mathscr{D}; 0, \Delta)$ in terms of x_l, x_{emp}, and S. First observe that

$$
y_l = x_l - \mu, \qquad y_{l-1} = \frac{l}{\sqrt{l-1}}(x_{\text{emp}} - \mu) - \frac{(x_l-\mu)}{\sqrt{l-1}},
$$

$$
\mathscr{D} = lS - \frac{l}{l-1}(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}. \tag{3.45}
$$

Taking into account that the Jacobian of the transformation (3.45) equals $l^{n(n+3)/2}/(l-1)^{n/2}$, and substituting (3.45) into (3.44), we obtain

$$q(x_l, x_{\text{emp}}, S; \mu, \Delta) = \frac{l^{n(n+3)/2}}{(l-1)^{n/2}} P\left(\frac{l}{\sqrt{l-1}}(x_{\text{emp}} - \mu) - \frac{(x_l - \mu)}{\sqrt{l-1}}; 0, \Delta\right)$$
$$\times P(x_l - \mu; 0, \Delta) W_{l-1}\left(lS - \frac{l}{l-1}(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}; \Delta\right),$$

whence

$$q(x_l, x_{\text{emp}}, S; \mu, \Delta)$$
$$= \begin{cases} \dfrac{l^{n(n+3)/2} C_{n,l-1} \left| lS - \dfrac{l(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right|^{(l-n-3)/2} |\mathscr{D}|^{l/2}}{(2\pi)^n (l-1)^{n(l-1)/2} |\Delta|^{l/2} \exp\left\{\dfrac{l}{2} \operatorname{Sp}[\Delta^{-1}(S + (x_{\text{emp}} - \mu)(x_{\text{emp}} - \mu)^{\mathrm{T}})]\right\}}, \\ \qquad \text{if } \left| S - \dfrac{(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right| > 0, \\ 0, \qquad \text{if } \left| S - \dfrac{(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right| = 0. \end{cases} \tag{3.46}$$

We shall now determine the denominator of the expression (3.43). For a normal distribution of vectors x, the statistics x_{emp} and lS are distributed independently:

$$P(x_{\text{emp}}, S; \mu, \Delta) = P(x_{\text{emp}}; \mu, \Delta) P(S; \Delta), \tag{3.47}$$

where x_{emp} is normally distributed with $N(\mu, (1/l)\Delta)$, and lS has the Wishart distribution $W_l(S; \Delta)$. This implies that

$$P(x_{\text{emp}}, S; \mu, \Delta) = \frac{C_{n,l}}{(2\pi)^{n/2}} \frac{l^{n/2} |S|^{(l-n-2)/2}}{|\Delta|^{l/2} \exp\left\{\dfrac{l}{2} \operatorname{Sp}[\Delta^{-1}(S + (x_{\text{emp}} - \mu)(x_{\text{emp}} - \mu)^{\mathrm{T}})]\right\}}, \tag{3.48}$$

if $|S| \geq 0$ and zero otherwise. $C_{n,l}$ is a constant defined in (3.31).

Substituting (3.46) and (3.48) into (3.43) we obtain

$$\varphi(x|t) = \frac{\Gamma\left(\dfrac{l-1}{2}\right)[(l-1)\pi]^{-n/2}}{\Gamma\left(\dfrac{l-n-1}{2}\right)|S|^{1/2}} \left(\frac{\left| S - \dfrac{(x - x_{\text{emp}})(x - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right|}{|S|} \right)^{(l-n-3)/2}$$

in the case when $|S| > 0$ and $|S - [(x - x_{\text{emp}})(x - x_{\text{emp}})^{\mathrm{T}}/(l-1)]| \geq 0$. Observe that

$$\frac{\left| S - \dfrac{(x - x_{\text{emp}})(x - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right|}{|S|} = \left(1 - \frac{(x - x_{\text{emp}})^{\mathrm{T}} S^{-1} (x - x_{\text{emp}})}{l-1}\right).$$

Hence we finally obtain

$$\varphi(x|x_{\text{emp}}, S) = \frac{\Gamma\left(\frac{l-1}{2}\right)}{[(l-1)\pi]^{n/2}\Gamma\left(\frac{l-n-1}{2}\right)|S|^{1/2}}\left[1 - \frac{(x-x_{\text{emp}})^{\mathrm{T}}S^{-1}(x-x_{\text{emp}})}{l-1}\right]_{+}^{(l-n-3)/2},$$

where

$$[z]_{+} = \begin{cases} z & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

§11 The Problem of Estimating the Parameters of a Density

It would thus seem that we have succeeded in achieving our goal of constructing a Bayesian estimator of a density and computing the best unbiased estimator. However, the methods which were utilized in obtaining these estimators substantially utilize special properties of the estimated density. Therefore the methods studied above are not the common ones for estimating densities of various types.

It is therefore of interest to study methods which perhaps do not yield such precise approximations as those studied above but which are regular, i.e., which could be used for estimating densities belonging to different parametric classes.

To obtain these methods we shall reformulate our problem. We shall assume that our purpose is the estimation of parameters of a density rather than density estimation. We also assume that if one succeeds in solving the intermediate problem of obtaining a "nice" estimator for the parameters of the density, then the density itself can be satisfactorily estimated by choosing as an approximation the density function $P(x, \alpha^*)$, where α^* are the estimated values of the parameters.

Observe that when the normal (Gaussian) distribution is estimated, neither the Bayes approximation nor the unbiased estimator of the density belongs to the class of normal distributions. In the case when the density is "assessed" by estimating its parameters, the approximations obtained belong to the Gaussian class. (This fact of itself is of no value. It only indirectly indicates how far the solution obtained may be from, say, the Bayesian one.)

Thus we shall estimate the parameters α_0 of the density $P(x, \alpha_0)$. The quantity

$$d(\alpha_0, \hat{\alpha}; x_1, \ldots, x_l) = |\alpha_0 - \hat{\alpha}(x_1, \ldots, x_l)|^2$$

will serve as the measure of the quality of the estimator $\hat{\alpha} = \hat{\alpha}(x_1, \ldots, x_l)$ of the vector of parameters $\alpha = \alpha_0$ based on the sample $x_1, \ldots, x_l$. The mathematical expectation of the quantity $d(\alpha_0, \hat{\alpha}; x_1, \ldots, x_l)$, i.e.,

$$d(\alpha_0, \hat{\alpha}, l) = \int d(\alpha_0, \hat{\alpha}; x_1, \ldots, x_l) P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l$$

serves as the measure of the quality of estimators of $\alpha = \alpha_0$ based on samples of size l (where $P(x_1, \ldots, x_l; \alpha_0)$ is the probability density of the sample $x_1, \ldots, x_l$).

Finally the quality of an estimator used for estimating the parameter α under the prior distribution $P(\alpha)$ will be measured by

$$R_B(\hat{\alpha}, l) = \int d(\alpha, \hat{\alpha}, l) P(\alpha)\, d\alpha. \tag{3.49}$$

The estimator $\hat{\alpha}$ which yields the minimum of the functional (3.49) is called a *Bayesian estimator of parameters.*

As in the case of density estimation, the prior distribution $P(\alpha)$ of parameters α is usually unknown; therefore, as before, the minimax criterion

$$R_{\text{mnx}}(\hat{\alpha}, l) = \sup_{\alpha} d(\alpha, \hat{\alpha}, l)$$

makes sense. The vector $\hat{\alpha}$ which yields the minimum of $R_{\text{mnx}}(\hat{\alpha}, l)$ forms the *minimax estimator of parameters.* However, the construction of a regular method for parameter estimation of a density is associated with the idea of the best unbiased estimation rather than with the Bayesian or minimax estimation.

Definition. We say that estimator $\hat{\alpha} = \hat{\alpha}(x_1, \ldots, x_l)$ is an *unbiased estimator* of the vector of parameters α_0 if

$$\int \hat{\alpha}(x_1, \ldots, x_l) P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l = \alpha_0.$$

Consider first the case when the probability density $P(x, \alpha_0)$ depends only on a scalar parameter α_0. Then for the class of unbiased estimators, the remarkable *Rao–Cramèr inequality* is valid:

$$\int (\alpha_0 - \hat{\alpha}(x_1, \ldots, x_l))^2 P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l \geq \frac{1}{I_\Phi}, \tag{3.50}$$

where

$$I_\Phi = -\int \frac{d^2 \ln P(x_1, \ldots, x_l; \alpha_0)}{d\alpha^2} P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l.$$

The quantity I_Φ is called *Fisher's information quantity*. For an independent sample it equals

$$I_\Phi = -l \int \frac{d^2 \ln P(x, \alpha_0)}{d\alpha^2} P(x, \alpha_0)\, dx.$$

A derivation of the Rao–Cramèr inequality is given in practically all modern texts in statistics (see, e.g., [35, 49, 58]). The meaning of this inequality is that the variance of an unbiased estimator (and this variance measures the precision of estimation in the case of unbiased estimators) is never less than the inverse of the Fisher's information quantity.

Thus the right-hand side of the inequality (3.50) determines the limiting precision of unbiased estimation of a parameter. An estimator for which the inequality (3.50) becomes an equality is called *efficient*. The problem is to obtain a regular method for constructing efficient estimators of parameters for various parametric classes of densities.

An inequality analogous to (3.50) may be obtained also for simultaneous unbiased estimation of several parameters. In this case the *Fisher information matrix I* whose elements are

$$I_{ij} = -\int \frac{\partial^2 \ln P(x_1, \ldots, x_l; \alpha_0)}{\partial \alpha_i\, \partial \alpha_j} P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l,$$

$$i, j = 1, 2, \ldots, n,$$

serves as an analog of the information quantity.

For an independent sample $x_1, \ldots, x_l$ the elements I_{ij} are equal to

$$I_{ij} = -l \int \frac{\partial^2 \ln P(x, \alpha)}{\partial \alpha_i\, \partial \alpha_j} dx.$$

Let the Fisher information matrix I be nonsingular, and let the estimators $\hat{\alpha}_1(x_1, \ldots, x_l), \ldots, \hat{\alpha}_n(x_1, \ldots, x_l)$ be unbiased estimators of the parameters $\alpha_1^0, \ldots, \alpha_n^0$. Consider for these estimators a covariance matrix B, i.e., a matrix with the elements

$$b_{ij} = M(\alpha_i^0 - \hat{\alpha}_i(x_1, \ldots, x_l))(\alpha_j^0 - \hat{\alpha}_j(x_1, \ldots, x_l)).$$

Then a multidimensional analog of the Rao–Cramèr inequality is the following assertion: for any vector z and any unbiased estimators $\hat{\alpha}_1(x_1, \ldots, x_l), \ldots, \hat{\alpha}_n(x_1, \ldots, x_l)$, the inequality

$$z^{\mathrm{T}} B z \geq z^{\mathrm{T}} I^{-1} z \tag{3.51}$$

is valid. The meaning of this inequality is as follows: let the quality of the joint estimator of n parameters $\alpha_1^0, \ldots, \alpha_n^0$ be determined by the square of weighted sums of deviation (with weights $z = (z, \ldots, z_n)^{\mathrm{T}}$, $z_i \geq 0$) over all the estimated parameters:

$$T(x_1, \ldots, x_l) = \left(\sum_{i=1}^{n} z_i(\alpha_i^0 - \hat{\alpha}_i(x_1, \ldots, x_l)) \right)^2.$$

Then the mathematical expectation of $T(x_1, \ldots, x_n)$ is bounded from below by the quantity $z^T I^{-1} z$. In other words, no matter how the quality of the joint unbiased estimation of n parameters is measured (i.e., for any weights z_i), the bound

$$MT(x_1, \ldots, x_l) \geq z^T I^{-1} z$$

is valid. In particular it follows from the inequality (3.51) that the variance of the estimator with respect to each parameter separately satisfies the inequality (3.50). Indeed, (3.50) is obtained from (3.51) for the specific vector $z = (0, \ldots, 0, 1, 0, \ldots, 0)^T$.

Estimation methods which yield equality in (3.51) for all z are called *jointly efficient*. When estimating several parameters our goal is to find jointly efficient estimators.

§12 The Maximum-Likelihood Method

Unfortunately there is no "regular" method to obtain efficient estimators of parameters of density based on a sample of a fixed size. There is only a method which allows us to construct asymptotically efficient estimators. This is the *maximum-likelihood method* developed by R. A. Fisher [58]. However, before considering this method we shall introduce several notions which are necessary for classifying estimators obtained from samples of large size.

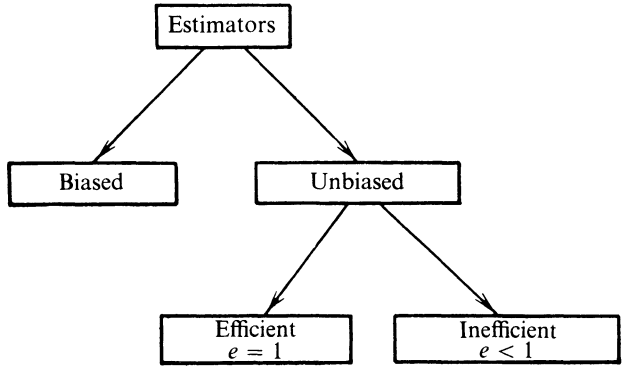

Figure 3

In the preceding section the classification presented here in Figure 3 was introduced for the characterization of estimators of parameters of a distribution based on samples of a fixed size. In this figure a measure of the efficiency of an unbiased estimator of parameters α_0 is also shown. In the

case of a single parameter this measure is given by

$$e_l = \frac{1}{M(\alpha_0 - \hat{\alpha}(x_1, \ldots, x_\alpha))^2 I_\Phi}. \tag{3.52}$$

In the case of joint estimation of several parameters the measure of efficiency is defined by

$$e_l = \frac{v(B, l)}{v(I, l)}, \tag{3.53}$$

which equals the ratio of the volume $v(B, l)$ of the ellipsoid

$$z^T B z = 1$$

to the volume of the ellipsoid

$$z^T I^{-1} z = 1.$$

For sample of large size a somewhat different classification is used which incorporates the notions of asymptotically unbiased, consistent, and asymptotically efficient estimators. Estimators satisfying

$$M_{\alpha_0} \hat{\alpha}(x_1, \ldots, x_l) \underset{l \to \infty}{\longrightarrow} \alpha_0$$

are called *asymptotically unbiased.* Estimators satisfying

$$P_{\alpha_0}\{|\hat{\alpha}(x_1, \ldots, x_l) - \alpha_0| > \varepsilon\} \underset{l \to \infty}{\longrightarrow} 0$$

for all $\varepsilon > 0$ are called *consistent.* Asymptotic unbiased estimators satisfying

$$e_l \underset{l \to \infty}{\longrightarrow} 1$$

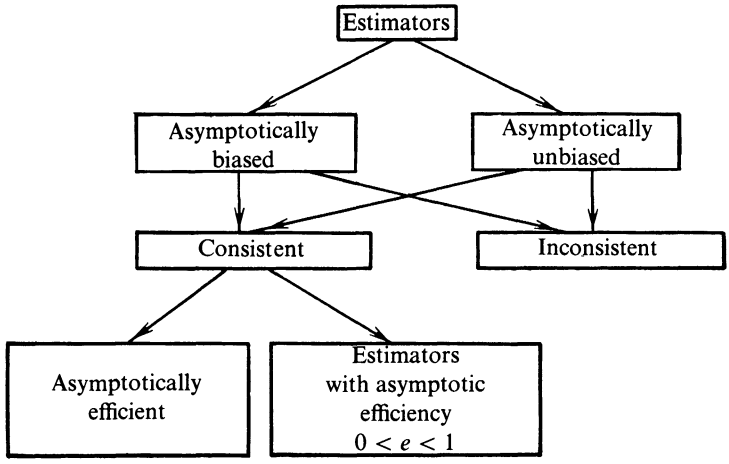

Figure 4

are called *asymptotically efficient*. Here e_l is given by (3.52) in the case of a single parameter α and by (3.53) when several parameters are jointly estimated. This classification is presented in Figure 4.

The method of maximum likelihood involves examining the likelihood function $P(x_1, \ldots, x_l; \alpha)$. In our case, when the sample $x_1, \ldots, x_l$ is obtained as a result of random independent observations according to the density $P(x, \alpha)$, the likelihood function can be represented as

$$P(x_1, \ldots, x_l; \alpha) = \prod_{i=1}^{l} P(x_i, \alpha). \tag{3.54}$$

The method of maximum likelihood chooses as the estimator those α which yield the maximum for (3.54). Along with the likelihood function (3.54) it is common to consider the function

$$\ln P(x_1, \ldots, x_l; \alpha) = \sum_{i=1}^{l} \ln P(x_i, \alpha). \tag{3.55}$$

The maxima of the functions (3.54) and (3.55) are the same, and hence to obtain maximum-likelihood estimators we need to solve the system of equations

$$\frac{\partial P(x_1, \ldots, x_l; \alpha)}{\partial \alpha_i} = 0, \qquad i = 1, 2, \ldots, n, \tag{3.56}$$

or the system of equations

$$\frac{\partial \ln P(x_1, \ldots, x_l; \alpha)}{\partial \alpha_i} = 0, \qquad i = 1, 2, \ldots, n. \tag{3.57}$$

The theory of maximum-likelihood estimation, which is well developed, aims to justify the applicability of this method. The substance of this theory is that for certain classes $P(x, \alpha)$ (to which all the classes of densities considered in this book belong) the maximum-likelihood method assures the asymptotic efficiency of the estimators (cf. [24, 58]).

We also remark that in the case of maximum-likelihood estimation the problem is reduced here to a simpler one than the one encountered in Bayesian estimation (multiple integration) or in constructing unbiased estimators (solution of Fredholm's equations of type I).

To implement the maximum-likelihood method it is necessary to solve the system of equations (3.56) or (3.57). Although this is not always a linear system, its numerical solution is not usually too difficult, and moreover, for a wide class of functions there exists a unique solution.

§13 Estimation of Parameters of the Probability Density Using the Maximum-Likelihood Method

In this section, utilizing the maximum-likelihood method, we shall obtain estimators for parameters of the distribution

$$P(x^i) = \begin{cases} p^i(1), & \text{for } x^i = c(1), \\ \vdots & \\ p^i(\tau_i), & \text{for } x^i = c(\tau_i), \end{cases} \qquad i = 1, 2, \ldots, n,$$

$$\sum_{j=1}^{\tau_j} p^i(j) = 1, \qquad i = 1, 2, \ldots, n,$$

as well as for parameters of the normal distribution

$$N(\mu, \Delta) = \frac{1}{(2\pi)^{n/2}|\Delta|^{1/2}} \exp\{-\tfrac{1}{2}(x - \mu)^{\mathrm{T}} \Delta^{-1}(x - \mu)\}.$$

It turns out that for the distribution $P(x^i)$ the estimators are given by

$$\hat{P}(x^i) = \begin{cases} \hat{p}^i(1) = \dfrac{m_i(1)}{l} & \text{for } x^i = c^i(1), \\ \vdots & \\ \hat{p}^i(\tau_i) = \dfrac{m_i(\tau_i)}{l} & \text{for } x^i = c^i(\tau_i), \end{cases} \tag{3.58}$$

where $m_i(j)$ is the number of vectors in the sample with the ith coordinate taking on the value $x^i = c^i(j)$.

Maximum-likelihood estimators of parameters of a multidimensional normal distribution are given by

$$x_{\text{emp}} = \frac{1}{l} \sum_{i=1}^{l} x_i,$$

$$S = \frac{1}{l} \sum_{i=1}^{l} (x_i - x_{\text{emp}})(x_i - x_{\text{emp}})^{\mathrm{T}}.$$

Thus we obtain the following estimator of the normal density:

$$\hat{P}(x) = \frac{1}{(2\pi)^{n/2}|S|^{1/2}} \exp\{-\tfrac{1}{2}(x - x_{\text{emp}})^{\mathrm{T}} S (x - x_{\text{emp}})\}. \tag{3.59}$$

13.1 Derivation in the Discrete Case

We estimate the parameters of the distribution $P(x^i)$. First we form the likelihood function:

$$P(x_1, \ldots, x_l; p) = \prod_{j=1}^{l} \prod_{i=1}^{n} P(x_j^i, p^i),$$

where x_j^i is the value of the ith coordinate of the j-vector in the sample.

Interchanging the order of the factors, we have

$$P(x_1, \ldots, x_l; p) = \prod_{i=1}^{n} \prod_{j=1}^{l} P(x_j^i, p^i).$$

We now proceed to the function

$$\ln P(x_1, \ldots, x_l; p) = \sum_{i=1}^{n} \sum_{j=1}^{l} \ln P(x_j^i, p^i).$$

Consider the quantity

$$\sum_{j=1}^{l} \ln P(x_j^i, p^i).$$

It can be represented in the form

$$\sum_{j=1}^{l} \ln P(x_j^i, p^i) = \sum_{r=1}^{\tau_i} m_i(r) \ln p^i(r),$$

where $m_i(r)$ is the number of vectors in the sample such that the ith coordinate takes the value $x^i = c^i(r)$. Thus

$$\ln P(x_1, \ldots, x_l; p) = \sum_{i=1}^{n} \sum_{r=1}^{\tau_i} m_i(r) \ln p^i(r). \tag{3.60}$$

We now obtain the maximum with respect to $p^i(r)$ of function (3.60) subject to $\sum_{r=1}^{\tau_i} p^i(r) = 1$, $i = 1, 2, \ldots, n$. For this purpose the method of Lagrange multipliers will be used. We form the Lagrange function

$$L(p, \lambda) = \sum_{i=1}^{n} \sum_{r=1}^{\tau_i} (m_i(r) \ln p^i(r) - \lambda_i p^i(r)), \tag{3.61}$$

where the λ_i are the Lagrange multipliers. The vector p^i which yields the maximum of $L(p, \lambda)$ is determined by the system of equations

$$\frac{\partial L(p^i, \lambda)}{\partial p^i(r)} = \frac{m_i(r)}{p^i(r)} - \lambda_i = 0, \qquad i = 1, \ldots, n. \tag{3.62}$$

From (3.62), taking the condition

$$\sum_{r=1}^{\tau_i} p^i(r) = 1,$$

into account, we obtain

$$\hat{p}^i(r) = \frac{m_i(r)}{l}.$$

Observe that here the maximum-likelihood estimators turn out to be unbiased.

13.2 Derivation in the Normal Case

We now estimate the parameters μ and Δ of the normal distribution:

$$P(x;\mu,\Delta)=\frac{1}{(2\pi)^{n/2}|\Delta|^{1/2}}\exp\{-\tfrac{1}{2}(x-\mu)^{\mathrm{T}}\Delta^{-1}(x-\mu)\}.$$

We form the likelihood function

$$P(x_1,\ldots,x_l;\mu,\mathscr{D})=\frac{|\mathscr{D}|^{l/2}}{(2\pi)^{ln/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{l}(x_i-\mu)^{\mathrm{T}}\mathscr{D}(x_i-\mu)\right\},$$

where $\Delta^{-1}=\mathscr{D}$. We obtain its logarithm

$$\ln P(x_1,\ldots,x_l;\mu,\mathscr{D})=-\frac{nl}{2}\ln 2\pi+\frac{l}{2}\ln|\mathscr{D}|-\frac{1}{2}\sum_{i=1}^{l}(x_i-\mu)^{\mathrm{T}}\mathscr{D}(x_i-\mu).$$

Write

$$\frac{\partial P(x_1,\ldots,x_l;\mu,\mathscr{D})}{\partial\mu}=\mathscr{D}\left(\sum_{i=1}^{l}x_i-l\mu\right)=0, \tag{3.63}$$

$$\frac{\partial P(x_1,\ldots,x_l;\mu,\mathscr{D})}{\partial\mathscr{D}}=\frac{l}{2}\mathscr{D}^{-1}-\frac{1}{2}\sum_{i=1}^{l}(x_i-\mu)(x_i-\mu)^{\mathrm{T}}=0. \tag{3.64}$$

Here we have used the relationship

$$\frac{d\ln|A|}{dA}=A^{-1}.$$

From Equations (3.63) and (3.64) we obtain

$$x_{\mathrm{emp}}=\frac{1}{l}\sum_{i=1}^{l}x_i,\qquad S=\mathscr{D}^{-1}=\frac{1}{l}\sum_{i=1}^{l}(x_i-x_{\mathrm{emp}})(x_i-x_{\mathrm{emp}})^{\mathrm{T}}.$$

The estimator of the covariance matrix is biased.

§14 Remarks on Various Methods for Density Estimation

Three types of estimation for densities defined up to parameters were considered in this chapter: Bayesian, best unbiased, and those obtained using the maximum-likelihood method. For our specific problems of estimating densities of two classes (3.58) and (3.59), all three estimators were obtained. Which one is preferable for use in practice, then—which one should be substituted into (3.7) to obtain decision rules in a pattern recognition problem?

Theoretically the Bayesian is undoubtedly the preferable one. This estimator optimizes a functional which defines the quality of the estimator

in a reasonable manner. However, in order to obtain a Bayes estimator the prior distribution of parameters of the density must be known, i.e., a distribution which determines how often in practice a particular density is estimated. Usually this information is not available.

In Sections 6 and 7 Bayesian estimators were obtained for prior distributions which on the one hand contain fairly indefinite information but on the other yield a maximal simplification of calculations. How much confidence should be given to a Bayesian estimator based on one prior distribution if in practice another distribution is implemented? Only a qualitative answer is available to this question. As the sample size increases, the effect of the prior information on the Bayesian estimator decreases. Thus the use of the Bayesian estimator is justified by the belief that in practice the inconsistency in the choice of a prior distribution has little effect.

When constructing the best unbiased estimator of a density there is no need to take prior information into account. In this class of estimators there exists a best estimator which is independent of a particular estimated density belonging to this class. It would seem that no risk is involved in choosing the best unbiased estimator in such a situation. Actually this is not the case. It does not follow at all that the class of unbiased estimators contains sufficiently "nice" estimators. It has already been mentioned that the unbiasedness by itself is of no value and is introduced only to restrict the class of estimators. The class of unbiased estimators is a narrow one (for example, an unbiased estimator of the normal distribution expressible in terms of sufficient statistics is unique). It is not excluded that the narrow class of unbiased estimators consists only of rather "inferior" estimators and then the choice of the best one in this class does not assure that the estimator is satisfactory.

The example suggested by C. Stein indicates that this indeed is quite possible: when estimating the mean vector μ of the n-dimensional ($n > 2$) normal distribution with unit covariance matrix I, the biased estimator

$$\hat{x}_{\text{emp}} = \left(1 - \frac{n-2}{l x_{\text{emp}}^{\mathrm{T}} x_{\text{emp}}}\right) x_{\text{emp}}$$

turns out to be a uniformly better estimator than the arithmetic mean

$$x_{\text{emp}} = \frac{1}{l} \sum_{i=1}^{l} x_i,$$

which is the best unbiased one. (More details on Stein-type estimators are given in Chapter 5.) Stein's example is remarkable in that it is constructed for the simplest problems of parameter estimation and even here uniformly better biased estimators exist.

Thus the choice of the best unbiased estimator can be justified only by the belief that the class of unbiased estimators contains an adequate one.

Finally, the theory of maximum-likelihood estimators provides no answers to the question concerning the properties of estimators for samples

of finite size. The theory only guarantees that the maximum-likelihood estimators approach the efficient ones as the sample size increases, i.e., with an increase in sample size, the quality of a maximum-likelihood estimator approaches that of the best unbiased estimator.

Due to a lucky contingency, we were able in this chapter to find Bayesian estimators explicitly, i.e., to carry out the analytic integration of a multiple integral (numerical integration of multiple integrals of high dimensions is troublesome) to obtain explicitly the best unbiased estimator of the density. That is, we were able to arrive at an analytic solution of Fredholm's Type I equation (whereas a numerical solution of this equation is an ill-posed problem). This result is due to a specific feature of the parametric class of densities.

In general, however, such approximations can hardly be anticipated. In this respect the maximum-likelihood method has an advantage in that it can be used for diverse classes of densities. This property of the maximum-likelihood method is due to the fact that it reduces to the solution of algebraic equations, i.e., to a problem for which efficient computer methods exist.

Yet another remark: The methods for estimating densities discussed in this chapter make sense only if the density under consideration belongs to a given parametric family of densities. In practice, however, the prior information which would allow us to select a parametric family of functions containing the unknown one is not available. It turns out, in fact, that not only the choice of a particular method of density estimation, but also the choice of a *parametric* formulation of the problem of estimating dependences from empirical data, is largely a matter of belief.

Chapter 4

Methods of Parametric Statistics for the Problem of Regression Estimation

§1 The Scheme for Interpreting the Results of Direct Experiments

In the preceding chapter methods of parametric statistics were applied to solve the pattern recognition problem: to minimize the functional

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \tag{4.1}$$

with unknown density $P(x, y)$, on the basis of empirical data

$$x_1, y_1; \ldots; x_l, y_l, \tag{4.2}$$

first the density $\hat{P}(x, y)$ was estimated in the parametric class of densities $\{P(x, y)\}$; then, using $\hat{P}(x, y)$, the empirical functional

$$I_{\text{emp}}(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(x, y)\, dx\, dy \tag{4.3}$$

was constructed; and finally a value α_{emp} was determined which minimizes (4.3).

To implement this process it was essential that the coordinate y take on only two values, zero and one; the set $F(x, \alpha)$ was a set of indicator functions, and the density $P(x, y)$ was a union of two densities. These were characteristic features of the pattern recognition problem. In this chapter we shall implement the same procedure of risk minimization, but in relation to the problem of regression estimation.

For a solution of this problem using methods of parametric statistics a specific model of density which differs from the one discussed in Chapter 3 is

used. It is assumed that the random variable y and a random vector x are related as follows:

$$y = F(x, \alpha_0) + \xi,$$

where $F(x, \alpha_0)$ is a function which belongs to the class $F(x, \alpha)$ and ξ is a noise independent of x distributed according to the density $P(\xi)$:

$$M\xi = 0, \qquad M\xi^2 < \infty.$$

Thus for any fixed x the distribution $P(\xi)$ induces the conditional density of y,

$$P(y|x) = P(y - F(x, \alpha_0)). \tag{4.4}$$

The joint density $P(x, y)$ is defined by

$$P(x, y) = P(y|x)P(x) = P(y - F(x, \alpha_0))P(x), \tag{4.5}$$

where $P(x)$ is the probability density of the vector x.

The problem of regression estimation, $F(x, \alpha_0) \in F(x, \alpha)$, based on a random and independent sample of pairs $x_1, y_1, \ldots, x_l, y_l$, can be interpreted as the estimation of the functional dependence $F(x, \alpha_0)$ in the class $F(x, \alpha)$ based on direct observations which are carried out subject to additive noise at l randomly chosen points. In Chapter 1 this problem was called "interpretation of results of direct experiments".

We shall solve this problem using methods of parametric statistics. First we estimate the density

$$\hat{P}(y|x) = \hat{P}(y - F(x, \alpha^*)),$$

and then we obtain the minimum point for the empirical functional

$$I_{\text{emp}}(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(y - F(x, \alpha^*))P(x)\, dx\, dy. \tag{4.6}$$

First we show that the minimum of the functional (4.6) is attained at $\alpha = \alpha^*$. We utilize the identity

$$\begin{aligned} I_{\text{emp}}(\alpha) &= \int (y - F(x, \alpha))^2 \hat{P}(y - F(x, \alpha^*))P(x)\, dx\, dy \\ &= \int (y - F(x, \alpha^*))^2 \hat{P}(y - F(x, \alpha^*))P(x)\, dx\, dy \\ &\quad + \int (F(x, \alpha) - F(x, \alpha^*))^2 P(x)\, dx. \end{aligned} \tag{4.7}$$

Since the first summand on the right-hand side does not depend on α, the minimum of $I_{\text{emp}}(\alpha)$ is attained if the second nonnegative summand vanishes, i.e., at $\alpha = \alpha^*$. Thus the value of the vector $\alpha = \alpha^*$ which defines the conditional density $\hat{P}(y|x) = \hat{P}(y - F(x, \alpha^*))$ immediately determines the regression. It equals $F(x, \alpha^*)$.

§2 A Remark on the Statement of the Problem of Interpreting the Results of Direct Experiments

In the statement of the problem of interpreting results of direct experiments it is required that the unknown function $F(x, \alpha_0)$ belong to a given parametric family $F(x, \alpha)$. This requirement is imposed because the density $P(y - F(x, \alpha))$ is to be estimated by methods of parametric statistics. However, another formulation is possible according to which the unknown density $P(x, y)$ belongs to a given parametric family of densities $P(x, y; \alpha)$ and the desired dependence $F(x, \alpha_0)$ does not belong to the given set of dependences $f(x, \beta)$. In other words, as the model for interpreting results of direct experiments the following problem may be posed: find the minimum of the functional

$$I(\beta) = \int (y - f(x, \beta))^2 P(y - F(x, \alpha_0)) P(x)\, dy\, dx \tag{4.8}$$

from the sample

$$x_1, y_1; \ldots; x_l, y_l$$

if the joint density $P(x, y) = P(y - F(x, \alpha_0))P(x)$ is unknown, $F(x, \alpha_0) \in F(x, \alpha)$, and the set of functions $f(x, \beta)$ does not necessarily coincide with $F(x, \alpha)$. If $F(x, \alpha_0) \notin f(x, \beta)$, the minimum of the functional (4.8) is attained at a function belonging to $f(x, \beta)$ which is closest to $F(x, \alpha_0)$. The proximity is measured here in the L_P^2 sense:

$$\rho_L(F, f) = \left(\int (F(x, \alpha_0) - f(x, \beta))^2 P(x)\, dx \right)^{1/2}.$$

If however $F(x, \alpha_0) \in f(x, \beta)$, then the minimum coincides with the regression. (This fact also follows immediately from (4.7).) Thus the regression yields an absolute minimum for the functional (4.8).

For a known density $P(x)$ the solution of the minimization problem for the functional (4.8) may also be carried out by means of the methods of parametric statistics: based on sample (4.2), the density $\hat{P}(y - F(x, \alpha))$ is estimated and then the empirical functional

$$I_{\text{emp}}(\beta) = \int (y - f(x, \beta))^2 \hat{P}(y - F(x, \alpha^*)) P(x)\, dx\, dy$$

is minimized.

Observe that for the problem of pattern recognition the search for a conditional minimum (in the class $f(x, \beta)$) of a functional, rather than the absolute one, was the subject matter of discriminant analysis. As it was pointed out in Section 2 of Chapter 3, the *raison d'être* for this formulation was based on the fact that the sample size is finite and hence the density is estimated only approximately; thus the lower guaranteed minimum for the value of the expected risk can be obtained for a function belonging to a narrower class. An analogous situation arises for the interpretation of results of direct experiments based on finite samples: due to imprecisions in density estimation, the higher guaranteed proximity to regression may be attained at a function belonging to a narrower class $f(x, \beta)$. Methods for contracting classes of desired dependences in order to achieve a lower guaranteed expected risk will be discussed in Chapter 8.

§3 Density Models

Thus in order to estimate regression—under the conditions of the model for interpreting the results of direct experiments—it is sufficient to estimate the density $P(y - F(x, \alpha_0))$ defined up to the value of parameter α. In view of the stipulated model, the parametric family of densities $P(y - F(x, \alpha))$ which contains the desired one is determined firstly by the given parametric family of functions $F(x, \alpha)$ containing the regression $F(x, \alpha_0)$, and secondly by the known probability density for the noise $P(\xi)$.

The assignment of a class of functions $F(x, \alpha)$ containing the regression is an informal step in the formulation of the problem. This class should be assigned *a priori.*

As far as the probability density of errors is concerned, here the choice is in principle arbitrary. However, in the practice of direct experimentation certain typical situations arise connected with common mechanisms which yield observational errors. These mechanisms have been investigated. The following three probability densities are of importance for interpreting results of direct experiments: the uniform density, normal density, and Laplace density.

The *uniform probability density* given by

$$P(\xi) = \begin{cases} \dfrac{1}{2\Delta} & \text{for } |\xi| \leq \Delta, \\ 0 & \text{for } |\xi| > \Delta \end{cases}$$

is used for roundoff errors. For example, let a value of a certain large quantity x be measured up to its integer value. Then the error ξ which arises from the roundoff to the closest integer is often assumed to be distributed according to the distribution

$$P(\xi) = \begin{cases} 1 & \text{for } |\xi| \leq 0.5, \\ 0 & \text{for } |\xi| > 0.5. \end{cases}$$

The *Normal (Gaussian) density* given by

$$P(\xi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ - \frac{\xi^2}{2\sigma^2} \right\}$$

is used to describe errors arising when repeated physical measurements are performed under identical conditions. These conditions determine the value of the variance σ^2. For example, errors resulting in measuring distances by means of a theodolite carried out under the same conditions (the same illumination, humidity, air temperature, degree of atmospheric pollution, etc.) are commonly described by the normal density.

The *Laplace density* given by

$$P(\xi) = \frac{1}{2\Delta} \exp\left\{ - \frac{|\xi|}{\Delta} \right\}$$

is used to describe errors occurring in physical experiments carried out under changing conditions. For example, if measurements of distances take place in unequal cloudiness, at different times, and under different pollution conditions, measurement errors are commonly described by a Laplace distribution.

Each density $P(\xi)$ generates its own parametric set of densities

$$P(y - F(x, \alpha)).$$

In this chapter only the maximum-likelihood method will be used for estimating the density in various parametric families. This method is chosen because its implementation presents no technical difficulties. It is well suited to all the parametric families of densities under consideration.

Thus we shall use the method of maximum likelihood for estimating parameters of the conditional density

$$P(y|x) = P(y - F(x, \alpha_0))$$

from the random independent sample

$$x_1, y_1; \ldots; x_l, y_l$$

distributed according to the density

$$P(x, y) = P(y - F(x, \alpha_0))P(x).$$

For this purpose we write the likelihood function

$$P(x_1, y_1, \ldots, x_l, y_l; \alpha) = \prod_{i=1}^{l} P(y_i - F(x_i, \alpha))P(x_i), \tag{4.9}$$

and then express it as a product of two factors:

$$P_1(\alpha) = \prod_{i=1}^{l} P(y_i - F(x_i, \alpha)), \tag{4.10}$$

which is the likelihood function for the conditional density, and

$$P_2 = \prod_{i=1}^{l} P(x_i).$$

Since the factor P_2 does not depend on the parameter α, (4.9) and (4.10) have the same maximum points. In what follows, the maximization of the function (4.10) will also be called a method of maximum likelihood.

We shall now consider the likelihood function $P_1(\alpha)$ for different distributions of the noise and find the corresponding maximum point.

The likelihood function (4.10) for the uniform distribution of ξ is of the form

$$P_1(\Delta, \alpha) = \prod_{i=1}^{l} \frac{1}{2\Delta} \delta_i(\alpha) = \frac{1}{(2\Delta)^l} \prod_{i=1}^{l} \delta_i(\alpha),$$

where

$$\delta_i(\alpha) = \begin{cases} 1 & \text{for } |y_i - F(x_i, \alpha)| \leq \Delta, \\ 0 & \text{for } |y_i - F(x_i, \alpha)| > \Delta. \end{cases}$$

The maximum of the likelihood function is determined by α and Δ for which the minimum of the expression

$$\Delta(\alpha) = \max_{x_i y_i} |y_i - F(x_i, \alpha)| \tag{4.11}$$

is attained, i.e., α is chosen to minimize the largest deviation of $F(x_i, \alpha)$ from y_i.

For the normal density the distribution of the likelihood function is given by the density

$$P_1(\alpha, \sigma) = \frac{1}{(2\pi)^{l/2}\sigma^l} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2 \right\},$$

and the maximum-likelihood method is equivalent to the minimization of the functional

$$I_{\text{emp}}(\alpha) = \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2. \tag{4.12}$$

The method of determining α by means of minimization of functional (4.12) is called the *least-squares method.*

Finally, if the error is distributed according to the Laplace density, then the corresponding likelihood function is

$$P_1(\Delta, \alpha) = \frac{1}{(2\Delta)^l} \exp\left\{ -\frac{1}{\Delta} \sum_{i=1}^{l} |y_i - F(x_i, \alpha)| \right\},$$

and the maximum of the likelihood is attained for the vector α for which the functional

$$I_{\text{emp}}(\alpha) = \sum_{i=1}^{l} |y_i - F(x_i \alpha)| \tag{4.13}$$

is minimized. The method of minimizing the functional (4.13) is called *the method of minimal modules.*

As was indicated in Chapter 3, the method of maximum likelihood is an asymptotically efficient method of estimating parameters; therefore all three algorithms are optimal in a certain sense. Unfortunately each one of them is optimal only under its own conditions (of uniform, normal, or Laplace distributions of errors), and solutions obtained by means of these algorithms may differ significantly.

Indeed, consider the simplest problem of estimating dependences—the determination of the mean value of a random variable y from a sample of size l. This problem is reduced to minimization of the functional

$$I(\alpha) = \int (y - \alpha)^2 P(y)\, dy \tag{4.14}$$

on the basis of a sample $y_1, \ldots, y_l$. Using the method of minimization of the largest deviation (4.11), we obtain the solution

$$\alpha^* = \frac{y_{\min} + y_{\max}}{2}, \tag{4.15}$$

where $y_{\min}$ is the smallest and $y_{\max}$ is the largest sample value; i.e., the estimator is the half range of the sample. The method of least squares (4.12) yields the estimator

$$\alpha^* = \frac{1}{l} \sum_{i=1}^{l} y_i; \tag{4.16}$$

i.e., the estimator is the sample arithmetic mean. Finally, the method of minimal modules (4.13) leads us to the following solution: order the observations according to their magnitude,

$$y_{i_1}, \ldots, y_{i_l},$$

and compute the estimator using the formula

$$\alpha^* = \begin{cases} y_{i_{k+1}} & \text{for } l = 2k + 1, \\ \dfrac{y_{i_k} + y_{i_{k+1}}}{2} & \text{for } l = 2k. \end{cases}$$

§4 Extremal Properties of Gaussian and Laplace Distributions

In the preceding section it was shown that algorithms for estimating regression obtained by methods of parametric statistics depend on the model adopted for the errors. Therefore it is necessary to be able to identify situations in which particular models are to be used. It was pointed out that the uniform distribution is used for describing errors resulting from rounding off, Gaussian distributions for measurement errors under identical conditions, and the Laplace law for measurements under changing conditions. It would be desirable to make these recommendations more precise.

In this section we shall establish certain remarkable properties for the Gaussian and Laplace distributions. We shall see that the Gaussian distribution possesses certain extremal properties under the absolute stability of measuring conditions, while the Laplace distribution possesses analogous extremal properties under "maximal instability" of measuring conditions.

Thus we shall show that among all continuous densities with a given variance, the normal distribution possesses the largest entropy. In other words, the normal distribution is a "noise" distribution in which the size of the measurement is undetermined to the largest possible extent.

We shall estimate the degree of uncertainty of measurements, in the case when errors are determined by the probability density $P(\xi)$, by means of

Shannon's entropy

$$H(P) = -\int_{-\infty}^{\infty} P(\xi) \ln P(\xi)\, d\xi. \tag{4.17}$$

We shall obtain a function $P(\xi)$ obeying the restrictions

$$P(\xi) \geq 0, \tag{4.18}$$

$$\int_{-\infty}^{\infty} P(\xi)\, d\xi = 1, \tag{4.19}$$

$$\int_{-\infty}^{\infty} \xi P(\xi)\, d\xi = 0, \tag{4.20}$$

$$\int_{-\infty}^{\infty} \xi^2 P(\xi)\, d\xi = \sigma^2, \tag{4.21}$$

for which the maximum of the entropy (4.17) is attained. Here the conditions (4.18), (4.19) follow from the definition of the density, (4.20) reflects the unbiasedness of the error, and (4.21) fixes the class of densities of a given variance.

This problem is solved using the standard method of Lagrange multipliers to take the conditions (4.19)–(4.21) into account:

$$L = -(P(\xi) \ln P(\xi) + \lambda_1 P(\xi) + \lambda_2 \xi P(\xi) + \lambda_3 \xi^2 P(\xi)).$$

We then write the Euler condition

$$\frac{\partial L}{\partial P} = -(\ln P(\xi) + 1 + \lambda_1 + \lambda_2 \xi + \lambda_3 \xi^2) = 0. \tag{4.22}$$

The solution of Equation (4.22),

$$P(x) = \exp\{-(\lambda_1 + 1 + \xi\lambda_2 + \xi^2\lambda_3)\},$$

satisfies (4.18) and hence determines the desired density.

To obtain values of the constants λ_1, λ_2, and λ_3 the conditions (4.19)–(4.21) are utilized; we obtain

$$P(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}, \tag{4.23}$$

thus the normal density has the largest entropy among all densities with a given variance (i.e., the random variable has the most "uncertain" distribution).

Consider now a somewhat more complicated model for the error term ξ. The value of random variable ξ is a realization of the normal distribution $P_N(\xi|\sigma^2)$ with mean 0 and variance σ^2. However, each time the normal distribution has its own variance. The value of the variance is assigned randomly and independently according to the density $P(\sigma^2)$. Thus we have

the distribution

$$P_\Lambda(x) = \int P_N(\xi \mid \sigma^2) P(\sigma^2)\, d\sigma^2. \tag{4.24}$$

This model reflects well the practical situation when under fixed conditions of measurements the normal distribution is valid. However, the measurement conditions change randomly and independently, and thus the probability density is a composition of two densities. In the example of measuring distances the factor $P_N(x \mid \sigma^2)$ in (4.24) reflects the errors occurring under the same atmospheric conditions. The factor $P(\sigma^2)$ reflects the random nature of the atmospheric conditions. If the measurement conditions do not change (the extreme case when $P(\sigma^2) = \delta(\sigma^2 - \sigma_0^2)$ where $\delta(z)$ is the delta function). then the composition (4.24) defines a normal distribution. We, however, shall consider the other extreme case when the experimental conditions deviate from the mean in the "most uncertain manner", i.e., when the function $P(\sigma^2)$ is such that the maximum of the entropy

$$H(P) = -\int_0^\infty P(\sigma^2) \ln P(\sigma^2)\, d\sigma^2 \tag{4.25}$$

is attained and moreover the conditions

$$P(\sigma^2) \geq 0, \tag{4.26}$$

$$\int P(\sigma^2)\, d\sigma^2 = 1, \tag{4.27}$$

$$\int_0^\infty \sigma^2 P(\sigma^2)\, d\sigma^2 = 2\Delta^2 \tag{4.28}$$

are satisfied. The conditions (4.26) and (4.27) follow from the definition of the probability density. The restriction (4.28) determines the average conditions of conducting the experiment.

We thus derive the maximum of the entropy (4.25) under the conditions (4.26)–(4.28). Writing the corresponding Lagrange function—which takes (4.27) and (4.28) into account

$$L = -(P(\sigma^2) \ln P(\sigma^2) + \lambda_1 P(\sigma^2) + \lambda_2 \sigma^2 P(\sigma^2)),$$

we obtain the Euler equation

$$\frac{\partial L}{\partial P} = -(\ln P(\sigma^2) + 1 + \lambda_1 + \lambda_2 \sigma^2) = 0. \tag{4.29}$$

The solution of Equation (4.29) is

$$P(\sigma^2) = \exp\{-(\lambda_1 + 1 + \lambda_2 \sigma^2)\}$$

which satisfies (4.26) and thus determines the desired density. To find the values of constants λ_1 and λ_2 we substitute solution (4.29) into (4.27) and

(4.28), whence $\lambda_1 + 1 = -\ln 2\Delta^2$ and $\lambda_2 = 1/2\Delta^2$. Thus the "most uncertain" conditions for conducting the experiment are given by density

$$P(\sigma^2) = \frac{1}{2\Delta^2} \exp\left\{-\frac{\sigma^2}{2\Delta^2}\right\}. \tag{4.30}$$

We shall show that the probability density $P_\Delta(\xi)$ given as a composition of densities (4.23) and (4.30) is a Laplace distribution, i.e.,

$$\begin{aligned} P_\Delta(\xi) &= \frac{1}{\sqrt{2\pi}2\Delta^2} \int_0^\infty \frac{1}{\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} \exp\left\{-\frac{\sigma^2}{2\Delta^2}\right\} d\sigma^2 \\ &= \frac{1}{2\Delta} \exp\left\{-\frac{|\xi|}{\Delta}\right\}. \end{aligned} \tag{4.31}$$

In order to compute the integral (4.31) we shall use the following fact, which is valid for any integrable function on $(-\infty, \infty)$:

$$\int_0^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right] dx = a \int_0^\infty f(y^2)\, dy \qquad (a, b > 0). \tag{4.32}$$

To prove this identity set $y = \dfrac{x}{a} - \dfrac{b}{x}$. Then

$$\begin{aligned} \int_{-\infty}^\infty f(y^2)\, dy &= \int_0^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right]\left(\frac{1}{a} + \frac{b}{x^2}\right) dx \\ &= \frac{1}{a} \int_0^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right] dx + b \int_0^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right] \frac{dx}{x^2}. \end{aligned}$$

Substituting the variable $x = -ab/t$ in the last integral, we arrive at

$$\frac{1}{a} \int_{-\infty}^0 f\left[\left(\frac{t}{a} - \frac{b}{t}\right)^2\right] dt.$$

Thus

$$\int_{-\infty}^\infty f(y^2)\, dy = \frac{1}{a} \int_{-\infty}^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right] dx.$$

Hence (since the integrand is even) we obtain the identity (4.32).

We now transform the left-hand side of (4.31):

$$\begin{aligned} P_\Delta(\xi) &= \frac{1}{2\sqrt{2\pi\Delta^2}} \int_0^\infty \frac{1}{\sigma} \exp\left\{-\left(\frac{\sigma^2}{2\Delta^2} + \frac{\xi^2}{2\sigma^2}\right)\right\} d\sigma^2 \\ &= \frac{1}{\sqrt{2\pi\Delta^2}} \exp\left\{-\frac{|\xi|}{\Delta}\right\} \int_0^\infty \exp\left\{-\frac{1}{2}\left(\frac{\sigma}{\Delta} - \frac{|\xi|}{\sigma}\right)^2\right\} d\sigma. \end{aligned} \tag{4.33}$$

From (4.33) in view of (4.32) we obtain

$$P_{\Lambda}(\xi) = \frac{1}{\Delta} \exp\left\{-\frac{|\xi|}{\Delta}\right\} \int_0^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy$$

$$= \frac{1}{2\Delta} \exp\left\{-\frac{|\xi|}{\Delta}\right\}. \tag{4.34}$$

In other words the composition (4.31) of a normal distribution and distribution (4.30) results in Laplace density (4.34).

Thus we have shown that under fixed conditions of conducting an experiment the most undetermined (uncertain) result is obtained when the error is normally distributed; if however the conditions of the experiment oscillate around some mean value in the most unfavorable manner, then the most undetermined measurement result is obtained when the error is distributed according to the Laplace law. Thus the choice between a Gaussian and a Laplace law depends on whether the conditions of the experiment are perfectly stable or most unstable.

In practice, however, these two extreme cases seldom occur. Therefore neither Gaussian nor Laplace distributions are usually fulfilled. It is customary to assume that an "intermediate" situation is valid.

Thus we are confronted with a situation where regression is estimated under the assumption that some hypothetical distribution for the error is valid (e.g., Gaussian or Laplace) while actually some other "intermediate" distribution is the correct one. How useful will the algorithms given by (4.11)–(4.13) then be? In other words, to what extent are the algorithms constructed robust as far as the changes in the distribution of errors are concerned, and how should one construct robust algorithms? The answer is given in the succeeding sections.

§5 On Robust Methods of Estimating Location Parameters

Let the probability density of the error be unknown. Suppose it is only known that it belongs to a certain given set of densities $\{P(\xi)\}$. Below we shall define such sets more precisely; for the time being we merely assume that they are convex and that the density functions possess two continuous derivatives and are symmetric around zero. (The symmetry is the basic requirement for the theory discussed below.) The following problem will now be investigated. How should one choose the hypothetical density for the noise from the given class $\{P(\xi)\}$ in order that the possible error shall have the least effect on the

estimators of regression parameters if it is known that the true density belongs to $\{P(\xi)\}$?

First consider the simple case: it is required to estimate the mathematical expectation m of a random variable x on the basis of the sample $x_1, \ldots, x_l$. If the mathematical expectation m exists the problem is equivalent to estimating the location parameter m of the density $P(x - m)$ (here we utilize the fact that the noise ξ is related to the measurement x by $\xi = x - m$). For a known density $P(\xi)$ the estimator $\hat{m}$ of location parameter m is carried out by the maximum-likelihood method, i.e., by maximizing the expression

$$R(m) = \sum_{i=1}^{l} \ln P(x_i - m). \tag{4.35}$$

In this case the estimator $\hat{m}$ is consistent and asymptotically efficient. However, if the function $P(\cdot)$ in (4.35) does not coincide with the density function of the noise $P(\xi)$, estimator $\hat{m}$ yielding the maximum of (4.35) will in general be neither consistent nor asymptotically efficient.

Denote the value $\hat{m}$ maximizing (4.35) under the assumption that $P(\xi) = P_\Gamma(\xi)$ by $\hat{m} = m(x_1, \ldots, x_l; P_\Gamma(\xi))$. We shall now determine how to measure the accuracy of parameter estimation if it is assumed that the noise is distributed according to the distribution $P_\Gamma(\xi) \in \{P(\xi)\}$ while actually the true distribution is $P_0(\xi) \in \{P(\xi)\}$.

It is natural to use the quantity

$$R(P_\Gamma(\xi); x_1, \ldots, x_l) = (\hat{m}(x_1, \ldots, x_l; P_\Gamma(\xi)) - m)^2$$

as the accuracy of the estimator $\hat{m}$ based on a sample $x_1, \ldots, x_l$, assuming that the noise is distributed according to the distribution $P_\Gamma(\xi)$. (This quantity is the square of the deviation of the obtained value of the parameter from the true one.) The accuracy of estimating a location parameter based on samples of size l is naturally measured by the mathematical expectation of the quantity $R(P_\Gamma(\xi); x_1, \ldots, x_l)$:

$$\begin{aligned} D(P_0, P_\Gamma) &= MR(P_\Gamma(\xi); x_1, \ldots, x_l) \\ &= \int (\hat{m}(x_1, \ldots, x_l; P_\Gamma(\xi)) - m)^2 P_0(x_1 - m) \cdots \\ &\quad \times P_0(x_l - m)\, dx_1 \cdots dx_l. \end{aligned} \tag{4.36}$$

The quantity $D(P_0, P_\Gamma)$ depends on two probability densities belonging to the same class $\{P(\xi)\}$: the hypothetical density $P_\Gamma(\xi)$ (according to which the estimator $\hat{m}$ was constructed) and the true density $P_0(\xi)$ (according to which the mean square deviation was computed).

Below we shall utilize the following representation of the function $D(P_0, P_\Gamma)$:

$$D(P_0, P_\Gamma) = \frac{1}{I} \frac{\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\left(\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' P_0(\xi)\, d\xi\right)^2}$$

$$= \frac{1}{I} \frac{\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\left(\int \frac{P'_\Gamma(\xi) P'_0(\xi)}{P_\Gamma(\xi)}\, d\xi\right)^2}. \tag{4.37}$$

We shall verify this representation by carrying out a not quite rigorous but intuitively appealing argument. A rigorous theory of robust estimation is presented in [88].

Without loss of generality it may be assumed that the true value of the location parameter m is zero. Denote

$$f(\xi) = \frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)} = (\ln P_\Gamma(\xi))'.$$

Then using the maximum-likelihood method, the estimator $\hat{m}$ of the parameter $m = 0$ is obtained from the condition

$$\left(\sum_{i=1}^{l} \ln P_\Gamma(x_i - \hat{m})\right)' = \sum_{i=1}^{l} f(x_i - \hat{m}) = 0.$$

We now utilize an approximation which is valid for large l and for the symmetric densities considered herein:

$$\sum_{i=1}^{l} f(x_i - \hat{m}) \approx \sum_{i=1}^{l} f(x_i) - \hat{m} \sum_{i=1}^{l} f'(x_i) = 0, \tag{4.38}$$

hence

$$\hat{m} = \frac{\sum_{i=1}^{l} f(x_i)}{\sum_{i=1}^{l} f'(x_i)}.$$

Let l be so large that

$$\hat{m} \approx \frac{\frac{1}{l} \sum_{i=1}^{l} f(x_i)}{\int f'(x) P_0(x)\, dx}.$$

(In the derivation of this relation it was assumed that the integral in the denominator exists. For this purpose it is sufficient that the functions $f'(x)$ be bounded. Below we shall consider only densities satisfying $|(\ln P(\xi))'| < \text{const}$).

Compute now $D(P_0, P_\Gamma) = M\hat{m}^2$:

$$D(P_0, P_\Gamma) = \int \hat{m}^2 P_0(x_1), \ldots, P_0(x_l)\, dx_1, \ldots, dx_l$$

$$= \frac{1}{l^2} \frac{1}{\left[\int f'(x) P_0(x)\, dx\right]^2} \int \sum_{i,j}^{l} f(x_i) f(x_j) \times P_0(x_1) \cdots P_0(x_l)\, dx_1 \cdots dx_l.$$

Since the densities $P_0(x)$, $P_\Gamma(x)$ are symmetric, we have

$$\int f(x_i) f(x_j) P_0(x_1) \cdots P_0(x_l)\, dx_1 \cdots dx_l = 0, \qquad i \neq j.$$

Thus we obtain for large l

$$D(P_0, P_\Gamma) = \frac{1}{l^2} \frac{\int \sum_{i=1}^{l} f^2(x_i) P_0(x_i)\, dx_i}{\left(\int f'(x) P_0(x)\, dx\right)^2} = \frac{1}{l} \frac{\int f^2(x) P_0(x)\, dx}{\left(\int f'(x) P_0(x)\, dx\right)^2}.$$

Finally, returning to the original notation we obtain representation (4.37).

We have thus determined a criterion of quality for estimators of location parameters given that the true density is $P_0(\xi)$ and the hypothetized one is $P_\Gamma(\xi)$. Our goal now is to choose a density $P_\Gamma(\xi)$ which minimizes $D(P_0, P_\Gamma)$. It is easy to show (see below) that if the density $P_0(\xi)$ were known, the minimum of $D(P_0, P_\Gamma)$ would be obtained at $P_\Gamma(\xi) = P_0(\xi)$.

The problem is to choose $P_\Gamma(\xi)$ if it is known only that $P_0(\xi)$ belongs to the class $\{P(\xi)\}$. As usual in such situations one of two approaches—the Bayesian or the minimax—is taken.

In the first case, it is assumed that the probability for each density in $\{P(\xi)\}$ to be the true one is known *a priori*, and the measure of quality of estimators is chosen to be the average (with respect to the measure $\mu(P)$) quality, i.e.,

$$D_B(P_\Gamma) = \int D(P_0, P_\Gamma)\, d\mu(P_0).$$

The minimax approach suggests that we choose as a measure of quality the quantity $D(P_0, P_\Gamma)$ evaluated for the least favorable density $P_0(\xi) \in \{P(\xi)\}$, i.e., to evaluate the quality from the condition

$$D_{\text{mnx}}(P_\Gamma) = \max_{P_0} D(P_0, P_\Gamma).$$

Since the construction of a solution optimal in the Bayes sense encounters substantial difficulties here, we shall study only minimax solutions below. Thus we shall judge the quality of an estimator of a location parameter, obtained by means of the hypothetized density $P_\Gamma(\xi)$, by the quantity

$$D_{\max}(P_\Gamma) = \max_{P_0} D(P_0, P_\Gamma) = \max_{P_0} \frac{\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{I\left(\int \frac{P'_\Gamma(\xi) P'_0(\xi)}{P_\Gamma(\xi)}\, d\xi\right)^2}, \tag{4.39}$$

and attempt to obtain a hypothetical density $P_\Gamma(\xi)$ minimizing (4.39).

Such a statement of the problem yields a game-theoretic interpretation. Let there be two players—nature and a statistician—who possess the same set of strategies (functions $\{P(\xi)\}$) but opposite goals. The first player (nature) attempts to select a strategy (i.e., assign a true density $P_0(\xi)$) which will maximize the losses of the second player, while the second chooses a strategy (hypothetized density $P_\Gamma(\xi)$) which minimizes his loss. The amount of loss is determined by the functional (4.39).

It is required to obtain the optimal strategy for the second player, i.e., to be able, for a given class of densities, to choose a hypothetized density that will guarantee the minimum losses for the least favorable true density. The density obtained will be called *robust in the class* $\{P(\xi)\}$, and the method of estimation of a location parameter obtained by applying the maximum-likelihood method to the density obtained is called the *method of robust estimation of a location parameter.*

An important fact in the theory of robust estimation of a location parameter is that the game with the loss function (4.39) possesses on the convex set $\{P(\xi)\}$ a saddle point, i.e.,

$$\max_{P_0 \in \{P(\xi)\}} \min_{P_\Gamma \in \{P(\xi)\}} D(P_0, P_\Gamma) = \min_{P_\Gamma \{P(\xi)\}} \max_{P_0 \in \{P(\xi)\}} D(P_0, P_\Gamma).$$

Using this fact one can obtain an optimal strategy against nature.

We now utilize the Cauchy–Schwarz inequality

$$\left(\int a(x)b(x)\, d\mu(x)\right)^2 \le \int a^2(x)\, d\mu(x) \int b^2(x)\, d\mu(x). \tag{4.40}$$

Using this inequality we rearrange the denominator of (4.37):

$$D(P_0, P_\Gamma) = \frac{1}{I} \frac{\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\left(\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)} \frac{P'_0(\xi)}{P_0(\xi)}\right) P_0(\xi)\, d\xi\right)^2} \ge \frac{1}{I \int \left(\frac{P'_0(\xi)}{P_0(\xi)}\right)^2 P_0(\xi)\, d\xi}. \tag{4.41}$$

Observe that for $P_\Gamma(\xi) = P_0(\xi)$ the equality

$$D(P_0, P_0) = \frac{1}{l \int \left(\frac{P_0'(\xi)}{P_0(\xi)}\right)^2 P_0(\xi)\, d\xi} \tag{4.42}$$

is valid. It follows from (4.41) and (4.42) that the minimum of (4.39) is attained if $P_\Gamma(\xi) = P_0(\xi)$, i.e., the optimal strategies of nature and the statistician result in the same density. To obtain this density it is necessary to maximize (4.42) over the class $\{P(\xi)\}$ or equivalently to obtain in the class $\{P(\xi)\}$ a density which will minimize the functional

$$I_\Phi(P) = l \int \left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\, d\xi.$$

Observe that the functional $I_\Phi(P)$ is the Fisher information quantity (cf. Chapter 3, Section 11).

In Sections 7 and 8 we shall obtain for various classes of probability densities those which minimize the Fisher information quantity and thus find robust estimators (within these classes) of a location parameter. In the next section we shall extend the result obtained here to the case of estimating regression parameters.

§6 Robust Estimation of Regression Parameters

Let it be required to estimate the regression. We shall assume that the class of functions in which the estimation is carried out and to which the regression belongs is represented in the form

$$F(x, \alpha) = \sum_{r=1}^{n} \alpha_r \varphi_r(x),$$

where $\varphi_r(x)$ is a system of linearly independent functions. As above, the true and the hypothetized densities of errors $P_0(\xi)$ and $P_\Gamma(\xi)$ belong to the convex class $\{P(\xi)\}$. The densities are symmetric around zero and have a bounded second logarithmic derivative.

To estimate regression parameters we shall use the maximum-likelihood method, i.e., we shall obtain the vector α which maximizes the expression

$$\ln P_\Gamma(x_1, y_1; \ldots; x_l, y_l; \alpha) = \sum_{i=1}^{l} \ln P_\Gamma\left(y_i - \sum_{r=1}^{n} \alpha_r \varphi_r(x_i)\right). \tag{4.43}$$

Let this vector be $\alpha = \alpha^*$. Consider the vector of deviations of the obtained values of regression parameters α^* from the actual ones α_0:

$$\bar{\alpha} = (\alpha_0 - \alpha^*).$$

Form the covariance matrix B:

$$B = M\bar{\alpha} \cdot \bar{\alpha}^{\mathrm{T}},$$

which determines the quality of estimation of the vector of parameters α (cf. Chapter 3, Section 11).

Below, analogously to (4.37), we shall obtain that for l sufficiently large the equality†

$$B = \frac{1}{l} \frac{\int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\left(\int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)' P_0(\xi)\, d\xi\right)^2} \|k_{ij}\|^{-1} \tag{4.44}$$

is valid, where

$$k_{ij} = \frac{1}{l} \sum_{t=1}^{l} \varphi_i(x_t)\varphi_j(x_t) \approx \int \varphi_i(x)\varphi_j(x) P_0(x)\, dx.$$

Thus the elements of matrix B are proportional to

$$D(P_0, P_\Gamma) = \frac{1}{l} \frac{\int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\left(\int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)' P_0(\xi)\, d\xi\right)^2}.$$

In the representation (4.44) it is important that only the proportionality coefficient $D(P_0, P_\Gamma)$ (and not the matrix $\|k_{ij}\|$) depends on the densities $P_0(\xi)$ and $P_\Gamma(\xi)$. Therefore two quadratic forms $z^{\mathrm{T}}B_1 z$ and $z^{\mathrm{T}}B_2 z$ with the same matrix $\|k_{ij}\|$ but different values of $D(P_0, P_\Gamma)$ correspond to two different hypothetized densities $P_\Gamma(\xi)$ and $\hat{P}_\Gamma(\xi)$. These forms satisfy one of two relations: either

$$z^{\mathrm{T}}B_1 z \geq z^{\mathrm{T}}B_2 z \quad \text{for any } z$$

or

$$z^{\mathrm{T}}B_1 z < z^{\mathrm{T}}B_2 z \quad \text{for any } z,$$

depending on whether $D(P_0, P_\Gamma)$ or $D(P_0, \hat{P}_\Gamma)$ is the largest. It was shown in Section 11 of Chapter 3 that the minimum of the quadratic form $z^{\mathrm{T}}Bz$ defines jointly efficient estimators of the parameters. Thus the value of the coefficient $D(P_0, P_\Gamma)$ determines the quality of estimation of the parameters of a linear regression: the smaller $D(P_0, P_\Gamma)$ is, the better is the quality.

This means that in the case of estimating regression parameters the problem of choosing a robust density leads to a game between nature and the statistician. It was shown in the preceding section that in this game the optimal strategy for the statistician is to choose a density belonging to the

† We assume additionally that the matrix $\|k_{ij}\|$ is not singular.

class of densities $\{P(\xi)\}$ which yields the minimum of Fisher's information quantity

$$I_{\Phi}(P) = I \int \left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\, d\xi. \tag{4.45}$$

Thus, in order to obtain the best hypothetical model for the error in the class $\{P(\xi)\}$ it is necessary to find a function belonging to this class which minimizes (4.45). This density will be used for the determination of regression parameters using the maximum-likelihood method.

It remains to derive the relation (4.44). It is obtained analogously to (4.37). Denote $f(\xi) = P'_{\Gamma}(\xi)/P_{\Gamma}(\xi)$. Then the maximum of the likelihood function (4.43) is attained at values of α which satisfy the equations

$$\sum_{i=1}^{l} f\left(\xi_i - \sum_{r=1}^{n} \bar{\alpha}_r \varphi_r(x_i)\right)\varphi_k(x_i) = 0, \qquad k = 1, 2, \ldots, n.$$

Utilizing the approximation (4.38), we have

$$\sum_{i=1}^{l} f\left(\xi_i - \sum_{r=1}^{n} \bar{\alpha}_r \varphi_r(x_i)\right)\varphi_k(x_i)$$

$$\approx \sum_{i=1}^{l} \left[f(\xi_i) - f'(\xi_i) \sum_{r=1}^{n} \bar{\alpha}_r \varphi_r(x_i)\right]\varphi_k(x_i) = 0.$$

Due to the independence of ξ_i and x_i we then obtain, for l sufficiently large,

$$\frac{1}{l}\sum_{i=1}^{l} f(\xi_i)\varphi_k(x_i) - \int f'(\xi)P_0(\xi)\, d\xi \sum_{i=1}^{l}\left(\sum_{r=1}^{n} \bar{\alpha}_r \varphi_r(x_i)\right)\varphi_k(x_i) = 0,$$

$$k = 1, 2, \ldots, n,$$

or in vector form,

$$\|k_{ij}\|\bar{\alpha} \approx \frac{1}{l}\frac{H}{\int f'(\xi)P_0(\xi)\, d\xi}, \tag{4.46}$$

where H is a column vector with coordinates $h_r = \sum_{i=1}^{l} \varphi_r(x_i) f(\xi_i)$.

It follows from (4.46) that

$$\bar{\alpha} = \frac{1}{l}\frac{1}{\int f'(\xi)P_0(\xi)\, d\xi} \|K_{ij}\|^{-1} H.$$

We now obtain the covariance matrix

$$B = M\bar{\alpha}\bar{\alpha}^{\mathrm{T}} = \frac{1}{l}\frac{\int f^2(\xi_i)P_0(\xi)\, d\xi}{\left(\int f'(\xi)P_0(\xi)\, d\xi\right)^2}\|k_{ij}\|^{-1}.$$

Returning to the original notation, we arrive at (4.44).

§7 Robustness of Gaussian and Laplace Distributions

We shall show that Gaussian and Laplace distributions are robust, each in its own class. As was shown in the preceding section, it is sufficient for this purpose to show that in corresponding classes of densities $\{P(\xi)\}$ the Gaussian and Laplace distributions yield the minimum of Fisher's information quantity (4.45).

For specific classes $\{P(\xi)\}$ which are discussed below this problem becomes a difficult problem in the calculus of variations (the class $\{P(\xi)\}$ is defined by restrictions of the inequality type). Therefore we shall not obtain the hypothetical density here by using a regular method, i.e., by solving nonclassical variational problems, but rather we shall first identify these solutions and then verify that they indeed define a saddle point of the function

$$D(P, P_\Gamma) = \frac{1}{I} \frac{\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P(\xi)\, d\xi}{\left(\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' P(\xi)\, d\xi\right)^2}.$$

In other words it will be required to verify that for a given density $P_\Gamma(\xi)$ the inequalities

$$D(P, P_\Gamma) \leq D(P_\Gamma, P_\Gamma) \leq D(P_\Gamma, P)$$

are fulfilled. Observe that in view of (4.41) one of the inequalities, namely

$$D(P_\Gamma, P_\Gamma) \leq D(P_\Gamma, P)$$

is always valid. Thus in order to prove the optimality of the selected strategy it is sufficient to establish the validity of the inequality

$$D(P, P_\Gamma) \leq D(P_\Gamma, P_\Gamma). \tag{4.47}$$

We consider the following classes of densities.

(1) *The class of densities with a bounded variance.* The corresponding variational problem is to minimize the functional (4.45) in the class of functions satisfying the conditions

$$\text{(1)} \qquad P(\xi) > 0,$$

$$\text{(2)} \qquad \int P(\xi)\, d\xi = 1,$$

$$\text{(3)} \qquad \int \xi P(\xi)\, d\xi = 0, \tag{4.48}$$

$$\text{(4)} \qquad \int \xi^2 P(\xi)\, d\xi \leq \sigma^2.$$

Conditions (1), (2), and (3) determine the density of the error term, and condition (4) is a bound on the variance. The solution of this nonclassical problem (in view of (1) and (4)) of the calculus of variations is the density

$$P(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}.$$

Indeed, substituting

$$P_\Gamma(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}$$

into the inequality (4.47), we obtain

$$\frac{\int \frac{\xi^2}{\sigma^4} P(\xi)\, d\xi}{\left(\frac{1}{\sigma^2}\int P(\xi)\, d\xi\right)^2} = \int \xi^2 P(\xi)\, d\xi \le \sigma^2. \tag{4.49}$$

This inequality is valid for any density belonging to (4.48), since the class (4.48) consists of densities for which the variance does not exceed σ^2. Thus the normal probability density with zero mean and variance σ^2 is robust in the class of all densities with the variance bounded by σ^2.

2. Now consider the *class of nondegenerate at zero densities.* Densities for which $P(0) \ge 1/2\Delta$ belong to this class. We shall show that the Laplace distribution is robust in this class of densities. For this purpose we substitute

$$P_\Gamma(\xi) = \frac{1}{2\Delta} \exp\left\{-\frac{|\xi|}{\Delta}\right\}$$

into (4.47). We obtain

$$\frac{\int \left(\frac{\operatorname{sign} \xi}{\Delta}\right)^2 P(\xi)\, d\xi}{\frac{4}{\Delta^2} P^2(0)} = \frac{1}{4P^2(0)} \le \Delta^2,$$

or equivalently

$$P(0) \ge \frac{1}{2\Delta}.$$

And since the densities satisfying $P(0) \ge 1/2\Delta$ are included in the class $\{P(\xi)\}$, the inequality (4.47) is satisfied for any function belonging to this class. Thus the Laplace distribution is robust in the class of densities for which $P(0) \ge 1/2\Delta$.

The robustness of the Gaussian and Laplace densities (each in its own class) is no less remarkable a fact than their extremal properties verified in Section 4.

Although the Gaussian and Laplace densities are robust, the class in which this property is valid often turns out to be exceedingly wide. In such cases a more meaningful statistical model should be constructed on the basis of other, narrower classes of densities.

Below in Sections 8 and 9 we shall consider certain specific classes of densities and obtain robust densities for these classes.

§8 Classes of Densities Formed by a Mixture of Densities

Consider the class H of densities formed by the mixture

$$P(\xi) = g(\xi)(1 - \varepsilon) + \varepsilon h(\xi) \tag{4.50}$$

of a certain fixed density $g(\xi)$ symmetric with respect to the origin and an arbitrary density $h(\xi)$ symmetric with respect to the origin. The weights in the mixture are $1 - \varepsilon$ and ε respectively. For classes of these densities the following theorem is valid.

Theorem 4.1 (Huber). *Let* $-\ln g(\xi)$ *be a twice continuously differentiable convex function. Then the class H possesses a robust density*

$$P_\Gamma(\xi) = \begin{cases} (1 - \varepsilon)g(\xi_0) \exp\{k(\xi - \xi_0)\}, & \text{for } \xi < \xi_0, \\ (1 - \varepsilon)g(\xi), & \text{for } \xi_0 \le \xi < \xi_1, \\ (1 - \varepsilon)g(\xi_1) \exp\{-k(\xi - \xi_1)\}, & \text{for } \xi \ge \xi_1, \end{cases} \tag{4.51}$$

where ξ_0 *and* ξ_1 *are the end points of the interval* $[\xi_0, \xi_1]$ *on which a monotone (due to the convexity of* $-\ln g(\xi)$*) function* $g'(\xi)/g(\xi)$ *is bounded in absolute value by a constant k determined by the normalization condition*

$$1 = (1 - \varepsilon) \int_{\xi_0}^{\xi_1} g(\xi)\, d\xi + \frac{g(\xi_0) + g(\xi_1)}{k}(1 - \varepsilon).$$

PROOF. To prove this theorem it is required to show (as in the case of proving robustness of Gaussian and Laplace densities) that functions belonging to the class (4.50) satisfy

$$D(P, P_\Gamma) \le D(P_\Gamma, P_\Gamma) \le D(P_\Gamma, P).$$

As has already been mentioned, the validity of the bound

$$D(P_\Gamma, P_\Gamma) \le D(P_\Gamma, P),$$

follows from the Cauchy–Schwarz inequality (4.40). Therefore to prove the theorem it is sufficient to verify that

$$D(P, P_\Gamma) \le D(P_\Gamma, P_\Gamma)$$

for any function $P(\xi) \in H$.

We represent the density $P_\Gamma(\xi)$ in the form of a mixture of a fixed density $g(\xi)$ and the density $\hat{h}(\xi) = [P_\Gamma(\xi) - (1-\varepsilon)g(\xi)]/\varepsilon$. We shall write the density $\hat{h}(\xi)$ explicitly taking (4.51) into account:

$$\hat{h}(\xi) = \begin{cases} \dfrac{1-\varepsilon}{\varepsilon}(g(\xi_0)\exp\{k(\xi-\xi_0)\} - g(\xi)) & \text{for } \xi < \xi_0, \\ 0 & \text{for } \xi_0 \le \xi < \xi_1, \\ \dfrac{1-\varepsilon}{\varepsilon}(g(\xi_1)\exp\{-k(\xi-\xi_1)\} - g(\xi)) & \text{for } \xi \ge \xi_1. \end{cases} \tag{4.52}$$

It is easy to verify that $\hat{h}(\xi)$ is a density. Indeed, $\int \hat{h}(\xi)\,d\xi = 1$, and $\hat{h}(\xi) \ge 0$, since by the assumption of the theorem $-\ln g(\xi)$ is a convex function and hence is situated totally above the tangent:

$$-\ln g(\xi) \ge -\ln g(\xi_i) - (-1)^i k(\xi - \xi_i), \qquad i = 0, 1. \tag{4.53}$$

This inequality is equivalent to the assertion

$$g(\xi) \le g(\xi_i)\exp\{(-1)^i k(\xi - \xi_i)\}, \qquad i = 0, 1.$$

Consider the inequality

$$\frac{\int\left(\dfrac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2[(1-\varepsilon)g(\xi) + \varepsilon h(\xi)]\,d\xi}{\left(\int\left(\dfrac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)'[(1-\varepsilon)g(\xi) + \varepsilon h(\xi)]\,d\xi\right)^2} \le \frac{(1-\varepsilon)\int\left(\dfrac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 g(\xi)\,d\xi + \varepsilon k^2}{(1-\varepsilon)^2\left(\int\left(\dfrac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' g(\xi)\,d\xi\right)^2}. \tag{4.54}$$

We shall verify that the right-hand side of this inequality is the least upper bound for the expression appearing in the left-hand side for arbitrary symmetric densities $h(\xi)$. For this purpose we observe that the function $P'_\Gamma(\xi)/P_\Gamma(\xi)$ equals

$$\frac{P'_\Gamma(\xi)}{P_L(\xi)} = \begin{cases} k & \text{for } \xi < \xi_0, \\ \dfrac{g'(\xi)}{g(\xi)} & \text{for } \xi_0 \le \xi < \xi_1, \\ -k & \text{for } \xi \ge \xi_1, \end{cases}$$

where according to the condition of the theorem $|g'(\xi)/g(\xi)| \le k$, and the function $(P'_\Gamma(\xi)/P_\Gamma(\xi))'$ equals

$$\left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' = \begin{cases} 0 & \text{for } \xi < \xi_0, \\ \left(\dfrac{g'(\xi)}{g(\xi)}\right)' & \text{for } \xi_0 \le \xi < \xi_1, \\ 0 & \text{for } \xi \ge \xi_1. \end{cases}$$

Thus in order to maximize the left-hand side of the inequality it is necessary to choose a density $h(\xi)$ which is situated on the intervals $(-\infty, \xi_0)$ and (ξ_1, ∞). Such a density simultaneously maximizes the numerator and minimizes the denominator of the expression appearing on the left-hand side of the inequality. The value of the expression appearing on the left will then be exactly equal to the value of the right-hand side of the inequality. The density (4.52) indeed belongs to the class of densities concentrated on the intervals $(-\infty, \xi_0)$, (ξ_1, ∞). The theorem is proved. □

This theorem is remarkable in that it allows us to construct various robust densities. In particular, if we choose for $g(\xi)$ the normal density

$$g(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\},$$

and consider the class of densities

$$P(\xi) = \frac{1-\varepsilon}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} + \varepsilon h(\xi),$$

then in view of the theorem the density

$$P_\Gamma(\xi) = \begin{cases} \dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma} \exp\left\{\dfrac{k^2}{2} - \dfrac{k}{\sigma}|\xi|\right\} & \text{for } |\xi| \geq k\sigma, \\ \dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma} \exp\left\{-\dfrac{\xi^2}{2\sigma^2}\right\} & \text{for } |\xi| < k\sigma \end{cases}$$

will be robust in this class, where k is determined from the normalization condition

$$1 = \frac{1-\varepsilon}{\sqrt{2\pi}\sigma}\left[\int_{-k\sigma}^{k\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} d\xi + \frac{2\exp\left\{-\dfrac{k^2}{2}\right\}}{k}\right].$$

The density just derived is an intermediate density between Gaussian and Laplace distributions. On the interval $|\xi| < k\sigma$ it coincides up to a normalizing constant with the Gaussian distribution and on the intervals $|\xi| \geq k\sigma$ with the Laplace distribution.

§9 Densities Concentrated on an Interval

We now consider yet another important class of densities and obtain a robust probability density in it.

Consider the class K_p *of densities concentrated on the whole on the interval* $[-A, A]$, i.e., the class of densities $P(\xi)$ for which the condition

$$\int_{-A}^{A} P(\xi)\, d\xi \geq 1 - p$$

is satisfied (where p is a known parameter which defines the class K_p). We shall show that in this class the density

$$P_\Gamma(\xi) = \begin{cases} \frac{1}{A}\left(\frac{b}{1+b}\cos^2\frac{a\xi}{A}\right) & \text{for } \left|\frac{\xi}{A}\right| < 1, \\ \frac{1}{A}\left(\frac{b}{1+b}\cos^2 a\right)\exp\left\{-2b\left(\left|\frac{\xi}{A}\right| - 1\right)\right\} & \text{for } \left|\frac{\xi}{A}\right| \geq 1 \end{cases} \tag{4.55}$$

is robust, where the parameters a, b are related to the constant p—which determines the class K—by the relations

$$p = 1 - \frac{\cos^2 a}{1+b},$$

$$b = a \tan a, \qquad 0 < a < \frac{\pi}{2}. \tag{4.56}$$

Without loss of generality it will be assumed that $A = 1$ (the class $A \neq 1$ is reduced to the case $A = 1$ by the substitution $z = A\xi$). Thus the problem is to show that in the class of densities satisfying the condition

$$\int_{-1}^{1} P(\xi)\, d\xi \geq 1 - p,$$

the density

$$P_\Gamma(\xi) = \begin{cases} \frac{b}{1+b}\cos^2 \xi\alpha & \text{for } |\xi| < 1, \\ \frac{b}{1+b}\cos^2 a \exp\{-2b(|\xi| - 1)\} & \text{for } |\xi| \geq 1 \end{cases} \tag{4.57}$$

will be robust. To do this it is sufficient to show that $P_\Gamma(\xi)$ given by (4.57) minimizes in K_p the Fisher functional

$$I_\Phi = I\int\left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\, d\xi. \tag{4.58}$$

Instead of directly minimizing the functional (4.58), however, we shall utilize the fact that the necessary and sufficient condition for $P_\Gamma(\xi)$ to be the minimum point for (4.58) is that the functional

$$R(P_\Gamma, P) = I\int(2(-\ln P_\Gamma(\xi))'' - [(\ln P_\Gamma(\xi))']^2)(P(\xi) - P_\Gamma(\xi))\, d\xi \tag{4.59}$$

is nonnegative in K_p. The functional $R(P_\Gamma, P)$ is the derivative with respect to ε of the expression

$$I_\Phi((1 - \varepsilon)P_\Gamma(\xi) + \varepsilon P(\xi)),$$

evaluated at $\varepsilon = 0$, i.e.,

$$\left.\frac{dI_\Phi((1-\varepsilon)P_\Gamma(\xi) + \varepsilon P(\xi))}{d\varepsilon}\right|_{\varepsilon=0} = R(P_\Gamma, P). \tag{4.60}$$

The nonnegativity of derivatives at $\varepsilon = 0$ (in any direction in K_p) for densities $(1-\varepsilon)P_\Gamma(\xi) + \varepsilon P(\xi)$ means that the minimum of I_Φ is attained on $P_\Gamma(\xi)$.

Thus we shall verify that the expression $R(P_\Gamma, P)$ is nonnegative. Since the function under the integral $R(P_\Gamma, P)$ is even, it is sufficient to verify that it is positive on the ray $0 \le \xi < \infty$. First note that (4.57) implies that

$$(-\ln P_\Gamma(\xi))' = \begin{cases} 2a \tan a\xi & \text{for } |\xi| < 1, \\ 2b \operatorname{sign} \xi & \text{for } |\xi| \ge 1. \end{cases} \tag{4.61}$$

Substituting (4.61) into (4.69) and carrying out the calculations, we have

$$R(P_\Gamma, P) = 4a^2 l \int_0^1 (P(\xi) - P_\Gamma(\xi))\, d\xi - 4b^2 l \int_1^\infty (P(\xi) - P_\Gamma(\xi))\, d\xi. \tag{4.62}$$

Transforming (4.62), we have

$$\begin{aligned} R(P_\Gamma, P) &= 4a^2 l \int_0^1 (P(\xi) - P_\Gamma(\xi))\, d\xi - 4b^2 l \int_1^\infty (P(\xi) - P_\Gamma(\xi))\, d\xi \\ &= 4(a^2 + b^2) l \int_0^1 (P(\xi) - P_\Gamma(\xi))\, d\xi. \end{aligned}$$

Thus the expression $R(P_\Gamma, P)$ is nonnegative for all $P(\xi)$ such that

$$\int_{-1}^1 P(\xi)\, d\xi \ge \int_{-1}^1 P_\Gamma(\xi)\, d\xi = 1 - 2\int_1^\infty P_\Gamma(\xi)\, d\xi = 1 - p,$$

i.e., for all functions belonging to K_p.

§10 Robust Methods for Regression Estimation

In preceding sections we have considered several classes of densities and obtained robust densities in these classes. It will now be possible in our scheme for interpreting results of direct experiments to weaken the requirements on prior information concerning the statistical properties of the errors. It is sufficient to know the class of densities to which the errors belong. In this case for estimating parameters of regression using methods of parametric statistics it is possible to use—instead of a true density—a density which is robust in the given class. Obviously this replacement reduces the asymptotic rate of convergence of parameters of the regression. This rate

becomes proportional to some quantity I situated in the interval

$$I_{\min} \leq I \leq I_{\max},$$

where

$$I_{\max} = \sup_{P(\xi) \in \{P(\xi)\}} \frac{1}{l \int \left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\, d\xi},$$

instead of being proportional to

$$I_{\min} = \frac{1}{l \int \left(\frac{P_0'(\xi)}{P_0(\xi)}\right)^2 P_0(\xi)\, d\xi},$$

which is the limiting value attainable in the case of unbiased estimation of the location parameter (cf. Chapter 3, Section 11) where $P_0(\xi)$ is the true density of the error. However, if the class $\{P(\xi)\}$ of densities is not too wide, then the possible loss of the rate is not overly large.

The basic constructive result of the theory of robust estimation considered here is the determination of four classes of densities with specified robust density.† We again identify these classes and their densities:

(1) *The class of densities with variance bounded by a constant* σ^2. A robust density in this class is the normal density

$$P_{\Gamma}(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}.$$

(2) *The class of nondegenerate densities* (for which $P(0) > 1/2\Delta$). In this class a robust density is

$$P_{\Gamma}(\xi) = \frac{1}{2\Delta} \exp\left\{-\frac{|\xi|}{\Delta}\right\}.$$

(3) *The class of densities formed by a mixture of a known density* (for example, a normal $P_N(\xi) = (1/\sqrt{2\pi}\sigma)e^{-\xi^2/2\sigma^2}$) with an arbitrary density in proportion $1 - \varepsilon : \varepsilon$. In this class the density

$$P_{\Gamma}(\xi) = \begin{cases} c \exp\left\{-\dfrac{\xi^2}{2\sigma^2}\right\} & \text{for } |\xi| < k\sigma, \\ c \exp\left\{\dfrac{k^2}{2} - k\left|\dfrac{\xi}{\sigma}\right|\right\} & \text{for } |\xi| \geq k\sigma \end{cases}$$

is robust (here c and k are constants determined by means of ε and σ).

† There are other classes of densities for which robust densities have been found [46].

(4) *The class of densities concentrated on the whole in the interval* $[-A, A]$ $\left(\int_{-A}^{A} P(\xi)\, d\xi \geq 1 - p\right)$. A density

$$P_\Gamma(\xi) = \begin{cases} c \cos^2 \dfrac{a\xi}{A} & \text{for } \left|\dfrac{\xi}{A}\right| < 1, \\ c \cos^2 a \exp\left\{-2b\left(\left|\dfrac{\xi}{A}\right| - 1\right)\right\} & \text{for } \left|\dfrac{\xi}{A}\right| \geq 1, \end{cases}$$

where c, a, and b are constants determined via A and p, is robust in this class.

Now suppose instead of the true density for the error $P_0(\xi)$ we choose a robust one in the class $P_\Gamma(\xi)$; determine, by means of it, the density of the conditional probability distribution

$$P_\Gamma\left(y - \sum_{r=1}^{n} \alpha_r \varphi_r(x)\right);$$

and finally utilize the maximum-likelihood method for parameter estimation. Then we arrive at the following algorithm of regression estimation based on the sample

$$x_1, y_1; \ldots; x_l, y_l.$$

One should minimize the functional

$$I_{\text{emp}}(\alpha) = \sum_{i=1}^{l} d\left(y_i - \sum_{r=1}^{n} \alpha_r \varphi_r(x_i)\right),$$

where

$$d(z) = z^2,$$

provided the true density of the error belongs to the class of densities with a bounded variance;

$$d(z) = |z|,$$

provided the true density of the error belongs to the class of nondegenerate densities;

$$d(z) = \begin{cases} \dfrac{z^2}{2\sigma^2} & \text{for } |z| < k\sigma, \\ -\dfrac{k^2}{2} + \dfrac{k}{\sigma}|z| & \text{for } |z| \geq k\sigma, \end{cases}$$

provided the true density is a mixture of a normal density with an arbitrary one;

$$d(z) = \begin{cases} -2 \ln \cos \dfrac{a}{A} z & \text{for } |z| < A, \\ b\left(\left|\dfrac{z}{a}\right| - 1\right) - 2 \ln \cos \dfrac{a}{A} z & \text{for } |z| \geq A, \end{cases}$$

provided the true density is concentrated on the whole on the interval $[-A, A]$.

Among these four methods, the least-square method ($d(z) = z^2$) and the method of minimal absolute values ($d(z) = |z|$) do not involve free parameters. The latter method is the most universal—it is determined by a stable density in a wider class of densities.

The other two methods of estimation involve parameters which are computed from the quantities defining the classes of densities. These methods should be used when possible to determine, as precisely as possible, the class of densities containing the desired one.

Thus when estimating regression we were able to remove the condition knowing exactly the error distribution. It is sufficient to know the class of functions which contains the regression and a class of densities to which the error density belongs. However, all of this theory developed for symmetric densities is essentially asymptotic (since in deriving the basic relation (4.37) the law of large numbers was substantially utilized). Therefore the belief that the asymptotic situation will occur rather early is the only guarantee that the algorithms obtained will be workable for samples of limited size.

Chapter 5

Estimation of Regression Parameters

§1 The Problem of Estimating Regression Parameters

In the previous section we considered methods for estimating regression under conditions when the sample size increases indefinitely. However, strictly speaking, the results were related to the problem of *estimating regression parameters* rather than the problem of *regression estimation.* This substitution (instead of approximating functions we estimate their parameters) is legitimate for samples of sufficiently large size. As the sample size increases, the estimated parameters approach the true values and hence the function constructed using these parameters tends to the regression function. However, for samples of limited size the estimation of the regression is not always equivalent to the estimation of its parameters.

Indeed, the quality of the estimator $\hat{\alpha}$ of the parameter α_0 of the regression $y(x) = F(x, \alpha_0)$ is determined by the proximity of the vectors α_0 and $\hat{\alpha}$:

$$\rho(\alpha_0, \hat{\alpha}) = \|\hat{\alpha} - \alpha_0\|, \tag{5.1}$$

whereas the quality of the approximation of a function $F(x, \hat{\alpha})$ to the regression $F(x, \alpha_0)$ is measured by the proximity of functions. In Chapter 1 we agreed to consider the mean-square measure of proximity

$$\rho_L(F(x, \alpha_0); F(x, \hat{\alpha})) = \left(\int (F(x, \hat{\alpha}) - F(x, \alpha_0))^2 P(x)\, dx\right)^{1/2}. \tag{5.2}$$

The criteria (5.1) and (5.2) are not identical, and it is possible that a solution which is the best according to one criterion may be the worst according to another.

EXAMPLE. In the class of functions

$$F(x, \alpha) = \alpha^0 + \alpha^1 x + \alpha^2 x^2$$

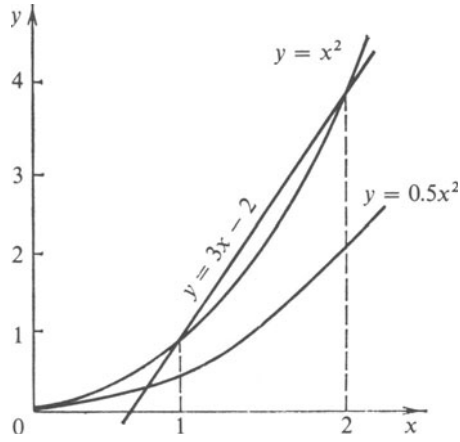

Figure 5

on the interval [1, 2], let the regression

$$y = x^2$$

be estimated. Consider two solutions (Figure 5): first the polynomial

$$F(x, \hat{\alpha}) = 0.5x^2$$

and second the polynomial

$$F(x, \hat{\hat{\alpha}}) = 3x - 2.$$

From the aspect of the parameter estimation criterion the first solution is better than the second (in any norm (5.1) the vector $\hat{\alpha} = (0, 0, 0.5)^T$ is closer to the vector $\alpha_0 = (0, 0, 1)^T$ than the vector $\hat{\hat{\alpha}} = (-2, 3, 0)^T$ is).

However, from the form of the criterion (5.2) the second solution $F(x, \hat{\hat{\alpha}})$ is better. For any measure $P(x)$ the inequality

$$\rho_L(3x - 2, x^2) < \rho_L(0.5x^2, x^2)$$

is valid.

When then is the problem of estimation of parameters of a regression based on samples of finite size equivalent to the problem of regression estimation?

Assume that the class of functions to which the regression belongs is linear in its parameters

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x), \tag{5.3}$$

and let $\varphi_1(x), \ldots, \varphi_n(x)$ be a system of orthonormal functions with weight $P(x)$, i.e., functions such that

$$\int_a^b \varphi_p(x)\varphi_q(x)P(x)\,dx = \begin{cases} 1 & \text{for } p = q, \\ 0 & \text{for } p \neq q. \end{cases} \tag{5.4}$$

In this case the quantities which characterize the proximity of functions in the L_P^2 metric and the proximity of parameters in the Euclidean metric coincide, and the problem of approximating a function on $[a, b]$ to the regression becomes equivalent to the problem of parameter estimation. Indeed,

$$\rho_L^2(F(x, \hat{\alpha}), F(x, \alpha)) = \int_a^b \left(\sum_{i=1}^n \hat{\alpha}_i \varphi_i(x) - \sum_{i=1}^n \alpha_i \varphi_i(x) \right)^2 P(x)\, dx$$

$$= \sum_{i=1}^n (\hat{\alpha}_i - \alpha_i)^2. \tag{5.5}$$

The conditions (5.3) and (5.4) are sufficient to replace the problem of estimating the regression with that of estimating its parameters. However, in order to construct an orthogonal system of functions the knowledge of $P(x)$ is needed. In this chapter we shall assume that the density $P(x)$ is known.

§2 The Theory of Normal Regression

The estimation theory of regression parameters based on samples of fixed size is developed for the case when the class of functions to which the regression belongs is linear in its parameters:

$$F(x, \alpha) = \sum_{i=1}^n \alpha_i \varphi_i(x), \tag{5.6}$$

and secondly the structure of the measurement follows the *Gauss–Markov model.* It is assumed that the measurements of functional dependence

$$y(x) = \sum_{i=1}^n \alpha_i^0 \varphi_i(x)$$

are carried out at l fixed points

$$x_1, \ldots, x_l.$$

(These points are not random.)

The measurements are subject to an additive noise which arises randomly according to the density $P(\xi)$, and has mean zero (i.e., $\int \xi P(\xi)\, d\xi = 0$) and finite variance ($\int \xi^2 P(\xi)\, d\xi < \infty$). The errors at points x_i and x_j ($i \neq j$) are uncorrelated.

The result of measurements of the function $\bar{y} = y(x)$ at points $x_1, \ldots, x_l$ is the random vector $Y = (y_1, \ldots, y_l)^T$ whose coordinates are equal to

$$y_j = \sum_{i=1}^n \alpha_i^0 \varphi_i(x_j) + \xi_j = \bar{y}_j + \xi_j, \qquad j = 1, 2, \ldots, l.$$

Using vector notation, we have

$$Y = \Phi\alpha_0 + \bar{\xi}, \tag{5.7}$$

where Φ is an $l \times n$ matrix with elements $\varphi_i(x_j)$ $(j = 1, 2, \ldots, l;\ i = 1, 2, \ldots, n)$, α_0 is the vector of parameters, and $\bar{\xi}$ is the noise vector. Thus the equalities

$$MY = \Phi\alpha_0, \qquad M\{(Y - MY)(Y - MY)^{\mathrm{T}}\} = \sigma^2 I, \tag{5.8}$$

where I is the unit matrix, define the Gauss–Markov model.

In the theory of estimating regression parameters, the special case of the Gauss–Markov model is considered for which the errors $\bar{\xi}$ are normally distributed.

For the normal distribution of the errors the so-called theory of *normal regression* is valid. It is based on the following fact: the extremal method of estimating parameters of normal regression is the least-squares method, according to which as an estimator of parameters α one should choose the vector α_{emp} which yields the minimum of the functional

$$I_{\text{emp}}(\alpha) = \frac{1}{l}\sum_{j=1}^{l}\left(y_j - \sum_{i=1}^{n}\alpha_i\varphi_i(x_j)\right)^2.$$

The following theorem is valid.

Theorem 5.1. *The least-squares estimators of parameters of a normal regression are jointly efficient.*

Below we shall prove this theorem and then construct a method estimating normal regression which is superior to the one based on the least-squares method.

PROOF. We write the probability density of the error in the form

$$P(\xi_j) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(y_j - \sum_{i=1}^{n}\alpha_i^0\varphi_i(x_j)\right)^2\right\}. \tag{5.9}$$

Here the problem of estimating regression parameters is equivalent to estimating the parameter of the distribution (5.9) based on the results of measuring the function $\bar{y} = y(x)$ at points $x_1, \ldots, x_l$.

We now write the likelihood function†

$$P(y_1, \ldots, y_l; \alpha) = P(\alpha)$$

$$= \frac{1}{(2\pi)^{l/2}\sigma^l}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{j=1}^{l}\left(y_j - \sum_{i=1}^{n}\alpha_i\varphi_i(x_j)\right)^2\right]\right\}. \tag{5.10}$$

† For brevity we shall write $P(\alpha)$ in place of $P(y_1, \ldots, y_l; \alpha)$.

In view of the Cramèr–Rao inequality (cf. Chapter 3, Section 11) the Fisher information matrix $\|f_{ij}\|$ (the matrix with elements

$$f_{ij} = -M\frac{\partial^2 \ln P(\alpha)}{\partial\alpha_i\,\partial\alpha_j}\Bigg)$$

determines the limiting accuracy of the joint estimators of the vector of parameters α in the class of unbiased estimators. Namely, for any vector z the inequality

$$z^{\mathrm{T}}\|f_{ij}\|^{-1}z \leq z^{\mathrm{T}}Bz$$

is valid, where B is the covariance matrix of unbiased estimators of the parameter vector. Thus the limiting accuracy in the class of unbiased estimators is attained for the estimation method for which

$$B = \|f_{ij}\|^{-1}. \tag{5.11}$$

We shall show that in the case of normal errors the equality (5.11) is attained when the regression parameters are estimated using the least-squares method. Indeed let us compute the elements f_{ij} of the Fisher matrix. Taking (5.10) into account we obtain

$$f_{ij} = -M\frac{\partial^2 \ln P(\alpha)}{\partial\alpha_i\,\partial\alpha_j} = \frac{1}{\sigma^2}M\sum_{r=1}^{l}\varphi_i(x_r)\varphi_j(x_r),$$

or in matrix form

$$\|f_{ij}\| = \frac{1}{\sigma^2}M\Phi^{\mathrm{T}}\Phi, \tag{5.12}$$

where Φ is an $l \times n$ matrix with elements $\varphi_i(x_j)$, $i = 1, \ldots, n, j = 1, \ldots, l$.

We now compute the elements b_{ij} of the covariance matrix B of estimators obtained using the least-squares method. For this purpose we shall find the estimator of regression parameters using the least-squares method, i.e., the vector α_{emp} which minimizes the functional

$$I_{\mathrm{emp}}(\alpha) = \frac{1}{l}\sum_{j=1}^{l}\left(y_j - \sum_{i=1}^{n}\alpha_i\varphi_i(x_j)\right)^2. \tag{5.13}$$

Minimization of $I_{\mathrm{emp}}(\alpha)$ with respect to α is equivalent to the solution of the following equation:

$$\Phi^{\mathrm{T}}\Phi\alpha = \Phi^{\mathrm{T}}Y. \tag{5.14}$$

Equation (5.14) is called the *normal equation.* A solution of the normal equation for the vector of parameters α equals†

$$\alpha = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}Y.$$

† It is assumed that $(\Phi^{\mathrm{T}}\Phi)$ is nonsingular; otherwise the generalized inverse $(\Phi^{\mathrm{T}}\Phi)^{+}$ is used in place of $(\Phi^{\mathrm{T}}\Phi)^{-1}$.

Observe that the least-squares estimator is unbiased:

$$M\alpha = M[(\Phi^T\Phi)^{-1}\Phi^T Y] = \alpha_0.$$

We now write the vector $\alpha - \alpha_0$ of deviations of estimators of regression parameters from the true value of parameters

$$\alpha - \alpha_0 = (\Phi^T\Phi)^{-1}\Phi^T\vec{\xi},$$

where $\vec{\xi}$ is the vector of errors in measurement.

Now we shall obtain the covariance matrix:

$$B = M(\alpha - \alpha_0)(\alpha - \alpha_0)^T = (\Phi^T\Phi)^{-1}\Phi^T M\vec{\xi}\vec{\xi}^T\Phi(\Phi^T\Phi)^{-1}.$$

Taking into account that $M\vec{\xi}\vec{\xi}^T = \sigma^2 I$, we arrive at

$$B = \sigma^2(\Phi^T\Phi)^{-1}.$$

Hence for the case of normally distributed errors the covariance matrix of vectors of estimators is equal to the inverse of the Fisher information matrix. We have thus shown the efficiency of the least-squares method for the problem of estimating regression parameters when the observations are assumed to follow the Gauss–Markov model. □

It should be mentioned that the least-squares method is an efficient method of estimating parameters only in the case of the Gauss–Markov model. In models with nonfixed measurement points x_i, even with normally distributed errors, the least-squares method is only asymptotically efficient. Thus even in the case of the estimation of one parameter,

$$\bar{y} = ax,$$

when measurements subject to additive normal error

$$y = ax + \xi$$

are taken at points $x_1, \ldots, x_l$ which are chosen randomly and independently according to distribution $P(x)$, the estimator of the parameter a is not efficient. Indeed, exactly as above one can find the value of the Fisher information quantity:

$$I_\Phi = \frac{M\sum_{i=1}^{l} x_i^2}{\sigma^2},$$

and compute the variance of the estimator of parameter a:

$$D(a) = M\frac{\sigma^2}{\sum_{i=1}^{l} x_i^2}.$$

Observe now that since the function $1/x^2$ is convex, the inequality

$$M \frac{1}{\sum_{i=1}^{l} x_i^2} \geq \frac{1}{M \sum_{i=1}^{l} x_i^2} \tag{5.15}$$

is valid. This implies that in the example under consideration

$$D(a) \geq I_{\Phi}^{-1}.$$

The only case when the inequality (5.15) becomes equality is when the observation points are fixed, which results in the Gauss–Markov model.

§3 Methods of Estimating the Normal Regression that are Uniformly Superior to the Least-Squares Method

Thus in the Gauss–Markov model the least-squares method is an efficient procedure for estimating parameters of a normal regression. This assertion required two stipulations:

(1) The observations are carried out with normal errors.
(2) The least-squares method is the best only among unbiased estimators.

The question arises: Are these stipulations essential? They are indeed. The least-squares method retains its extremal properties only in the case of normal errors ξ. When the number of observations $l \geq 2n + 1$ (n is the dimensionality of the basis), then the efficiency of the least-squares method implies that the errors are normally distributed [23].

No less important is the second stipulation: even under the conditions of normally distributed errors in a class of biased estimators, there exist estimators which are uniformly superior to the least-squares estimators.

Definition. We say that for the loss function

$$\|\alpha - \alpha_0\|^2 = (\alpha - \alpha_0)^{\mathrm{T}}(\alpha - \alpha_0),$$

the estimation method $\alpha_A(y_1, \ldots, y_l)$ of a vector of parameters α_0 is *uniformly better* than the estimation method $\alpha_B(y_1, \ldots, y_l)$ if for any α_0 the inequalities

$$M\|\alpha_A(y_1, \ldots, y_l) - \alpha_0\|^2 < M\|\alpha_B(y_1, \ldots, y_l) - \alpha_0\|^2$$

are satisfied.

In this section we shall construct algorithms for approximating regression which are uniformly better (i.e., better for any α_0) than those which result from the least-squares method. The bases for these algorithms are methods of

estimating the mean vector of a multivariate normal distribution, and in particular the following

Theorem 5.2 (James–Stein). *Let x be an n-dimensional ($n \geq 3$) random vector distributed according to a normal distribution $N(\alpha, \sigma^2 I)$ with the mean vector α and covariance matrix $\sigma^2 I$. Let S be a random variable independent of x distributed according to the central $\sigma^2\chi^2$ distribution with q degrees of freedom. Then the estimator of the mean given by*

$$\hat{\alpha}(x, S) = \left(1 - \frac{n-2}{q+2}\frac{S}{\|x\|^2}\right)_+ x, \qquad (z)_+ = \begin{cases} z & \text{for } z \geq 0, \\ 0 & \text{for } z < 0 \end{cases} \tag{5.16}$$

is uniformly better than $\hat{\hat{\alpha}}(x) = x$.

In other words, the theorem asserts that the vector $\hat{\alpha}(x, S)$ collinear to the observed vector x but different from x in its absolute value should be chosen as the estimator of α. This theorem is a particular case of a more general assertion to be proven in the next section.

We shall now utilize Theorem 5.2 to construct an algorithm for estimating regression which is uniformly superior to the one based on the least-squares method. Let observations $y_1, \ldots, y_l$ be carried out at the points $x_1, \ldots, x_l$; our purpose is to construct an approximation of a normal regression superior to the least-squares one. As above, we shall define proximity of functions using the L_P^2 metric:

$$\rho_L(F(x, \hat{\alpha}), F(x, \alpha)) = \left(\int (F(x, \hat{\alpha}) - F(x, \alpha))^2 P(x)\, dx\right)^{1/2}.$$

We now proceed to a doubly orthogonal basis

$$\psi_1(x), \ldots, \psi_n(x), \tag{5.17}$$

i.e., a basis which satisfies

$$\int \psi_i(x)\psi_j(x)P(x)\, dx = \begin{cases} \lambda_i & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \qquad \sum_{r=1}^{l} \psi_i(x_r)\psi_j(x_r) = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \tag{5.18}$$

and seek the regression expanded with respect to the basis (5.17)†

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \psi_i(x).$$

† According to the theorem on simultaneous reduction of two quadratic forms to a diagonal form using a linear transformation, such a basis exists and may be constructed using linear algebra.

In the new basis the proximity of the function $F(x, \alpha)$ to the regression $F(x, \alpha_0)$ is given by

$$\rho_L^2(F(x, \alpha), F(x, \alpha_0)) \equiv \rho_L^2(\alpha, \alpha_0) = \int \left(\sum_{i=1}^{n} (\alpha_i^0 - \alpha_i)\psi_i(x) \right)^2 P(x)\, dx = \sum_{i=1}^{n} \lambda_i (\alpha_i^0 - \alpha_i)^2.$$

Thus our purpose is to obtain an algorithm $\hat{\alpha}(y_1, \ldots, y_l)$ for estimating the parameter α_0 such that the quantity

$$M\rho_L^2(\hat{\alpha}(y_1, \ldots, y_l), \alpha_0) = M \sum_{i=1}^{n} \lambda_i (\hat{\alpha}_i(y_1, \ldots, y_l) - \alpha_i^0)^2 \tag{5.19}$$

is less than

$$M\rho_L^2(\alpha_{\text{lse}}, \alpha_0) = M \sum_{i=1}^{n} \lambda_i (\alpha_{\text{lse}}^i - \alpha_i^0)^2,$$

where $\alpha_{\text{lse}} = (\alpha_{\text{lse}}^1, \ldots, \alpha_{\text{lse}}^n)^{\mathrm{T}}$ is the least-squares estimator.

Consider now the least-squares estimator of regression parameters. In the basis (5.17) this estimator becomes

$$\alpha_{\text{lse}} = \Phi^{\mathrm{T}} Y,$$

where Φ is an $l \times n$ matrix with elements $\psi_i(x_j)$, $j = 1, \ldots, l$, $i = 1, \ldots, n$, and Y is the vector of observations. The vector α_{lse} is a random vector normally distributed with the mean vector

$$M\alpha_{\text{lse}} = M\Phi^{\mathrm{T}} Y = \alpha_0$$

and the covariance matrix $\sigma^2 I$:

$$M(\alpha_{\text{lse}} - \alpha_0)(\alpha_{\text{lse}} - \alpha_0)^{\mathrm{T}} = M\Phi^{\mathrm{T}} \bar{\xi} \bar{\xi}^{\mathrm{T}} \Phi = \sigma^2 I.$$

Thus the problem of estimating the parameter α_0 of the regression is reduced to the estimation of the mean vector α_0 of a normal distribution $N(\alpha_0, \sigma^2 I)$ based on its realization α_{lse}.

If in (5.19) all the λ_i were equal, Theorem 5.2 could be used to construct an algorithm for estimating regression which is better than the least-squares one. Indeed, as will be shown below, the statistic

$$S = Y^{\mathrm{T}} Y - \alpha_{\text{lse}}^{\mathrm{T}} \alpha_{\text{lse}} \tag{5.20}$$

does not depend on α_{lse} and is distributed according to the central $\sigma^2 \chi^2$ distribution with $l - n$ degrees of freedom. Therefore according to Theorem 5.2 the estimator

$$\hat{\alpha} = \left(1 - \frac{n-2}{l-n+2} \frac{Y^{\mathrm{T}} Y - \alpha_{\text{lse}}^{\mathrm{T}} \alpha_{\text{lse}}}{\alpha_{\text{lse}}^{\mathrm{T}} \alpha_{\text{lse}}} \right)_+ \alpha_{\text{lse}} \tag{5.21}$$

is uniformly better than α_{lse}, i.e., yields a value of the criterion (5.19) (in the case when $\lambda_1 = \cdots \lambda_n$) smaller than α_{lse}. However, in the doubly orthogonal

system (5.17) constructed above, not all λ_i are generally equal. Thus obtaining a better approximation to the regression in the case of unequal λ_i involves the determination of an estimation method yielding a value for the criterion (5.19) which is lower than that due to the least-squares method.

Construction of such an estimating algorithm is also based on the results of Theorem 5.2. We shall assume that the functions ψ_i are enumerated in increasing order of λ_i ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$). We shall introduce the following notation: let $\alpha_0(p)$ be a vector of dimensionality p, consisting of the first p coordinates of the vector $\alpha_0 = (\alpha_1^0, \ldots, \alpha_n^0)^{\mathrm{T}}$; let $\alpha_{\mathrm{lse}}(p)$ be the vector consisting of the first p coordinates of the vector of estimators obtained by the least-squares method α_{lse}.

Define n numbers $f_1, \ldots, f_n$:

$$f_1 = 1,$$

$$f_p = \left(1 - \frac{S}{\alpha_{\mathrm{lse}}^{\mathrm{T}}(p)\,\alpha_{\mathrm{lse}}(p)} \frac{p-2}{l-p+2}\right)_+, \qquad p = 2, \ldots, n.$$

Using these numbers, we construct n numbers h_p by the rule

$$h_p = \frac{\sum_{i=p}^{n} (\lambda_i - \lambda_{i+1}) f_i}{\lambda_p}, \quad \text{where } \lambda_{n+1} = 0, \qquad p = 1, 2, \ldots, n.$$

The following theorem is valid.

Theorem 5.3 (Bhattacharya). *For the risk function* (5.19) *the estimator*

$$\hat{\alpha}(y_1, \ldots, y_l) = (\alpha_{\mathrm{lse}}^1 h_1, \ldots, \alpha_{\mathrm{lse}}^n h_n)^{\mathrm{T}}, \qquad n \geq 3, \tag{5.22}$$

is uniformly better than the estimator $\alpha_{\mathrm{lse}} = (\alpha_{\mathrm{lse}}^1, \ldots, \alpha_{\mathrm{lse}}^n)^{\mathrm{T}}$.

PROOF. The proof of Theorem 5.3 is based on Theorem 5.2, according to which for any p the inequality

$$M\|\alpha_{\mathrm{lse}}(p) f_p - \alpha_0(p)\|^2 \leq M\|\alpha_{\mathrm{lse}}(p) - \alpha_0(p)\|^2 \tag{5.23}$$

is valid.

Consider the randomized estimator

$$g\alpha_{\mathrm{lse}} = (\alpha_{\mathrm{lse}}^1 g_1, \ldots, \alpha_{\mathrm{lse}}^n g_n), \tag{5.24}$$

where g_k are random variables independent of S and y distributed according to

$$P\{(q_k = f_j)\} = \frac{\lambda_j - \lambda_{j+1}}{\lambda_k}, \quad k = 1, 2, \ldots, n, \quad j = k, \ldots, n; \qquad \lambda_{n+1} = 0.$$

The value of this risk (5.19) for this estimator is equal to

$$\rho_L^2(G\alpha_{\text{lse}}, \alpha_0) = M\sum_{k=1}^{n} \lambda_k (g_k \alpha_{\text{lse}}^k - \alpha_k^0)^2$$

$$= \sum_{k=1}^{n} \sum_{j=k}^{n} \frac{\lambda_j - \lambda_{j+1}}{\lambda_k} \lambda_k M(f_j \alpha_{\text{lse}}^k - \alpha_k^0)^2.$$

We now utilize the inequality (5.23):

$$\rho_L^2(G\alpha_{\text{lse}}, \alpha_0) = \sum_{k=1}^{n} \sum_{j=k}^{n} (\lambda_j - \lambda_{j+1}) M(\alpha_{\text{lse}}^k f_j - \alpha_k^0)^2$$

$$= \sum_{j=1}^{n} (\lambda_j - \lambda_{j+1}) M \sum_{k=1}^{j} (\alpha_{\text{lse}}^k f_j - \alpha_k^0)^2$$

$$= \sum_{j=1}^{n} (\lambda_j - \lambda_{j+1}) M \|\alpha_{\text{lse}}(j) f_j - \alpha_0(j)\|^2$$

$$\leq \sum_{j=1}^{n} (\lambda_j - \lambda_{j+1}) M \|\alpha_{\text{lse}}(j) - \alpha_0(j)\|^2$$

$$\leq \sum_{j=1}^{n} \lambda_j M(\alpha_{\text{lse}}^j - \alpha_j^0)^2.$$

Thus the value of the risk for the randomized estimator of the parameters is less than the corresponding value for the least-squares estimator. On the other hand, it follows from the convexity of the loss function (5.19) that the nonrandomized estimator (5.22) is at least as good as the randomized estimator (5.24). Thus the approximation to the regression determined by the parameters (5.22) is uniformly better than the least-squares approximation. The theorem is proved. □

It remains to show that statistics $S = Y^{\mathrm{T}}Y - \alpha_{\text{lse}}^{\mathrm{T}} \alpha_{\text{lse}}$ does not depend on α_{lse} and is distributed according to the central $\sigma^2\chi^2$ distribution with $l - n$ degrees of freedom. For this purpose we shall complete the system of n vectors $\psi_1, \ldots, \psi_n$, orthonormal on $x_1, \ldots, x_l$:

$$\psi_i = (\psi_i(x_1), \ldots, \psi_i(x_l))^{\mathrm{T}},$$

$$\psi_i^{\mathrm{T}} \psi_j = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \qquad i, j = 1, 2, \ldots, n,$$

so that it becomes a complete orthonormal system consisting of l orthonormal vectors

$$\psi_1, \ldots, \psi_n, \psi_{n+1}, \ldots, \psi_l,$$

$$\psi_i^{\mathrm{T}} \psi_j = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \qquad i, j = 1, 2, \ldots, l.$$

We now expand Y in terms of this system:

$$Y = \sum_{i=1}^{n} \gamma_i \psi_i + \sum_{j=n+1}^{l} \gamma_j \psi_j, \tag{5.25}$$

where

$$\gamma_i = Y^{\mathrm{T}} \psi_i = \alpha_{\mathrm{lse}}^i, \qquad i = 1, 2, \ldots, n,$$

$$\gamma_j = Y^{\mathrm{T}} \psi_j, \qquad j = n+1, \ldots, l.$$

Substituting (5.25) into (5.20), we obtain

$$S = \sum_{j=n+1}^{l} \gamma_j^2, \tag{5.26}$$

and hence S does not depend on α_{lse}^i (but only on $\gamma_j, j = n+1, \ldots, l$). Since by assumption $Y = Y_0 + \bar{\xi}$ and the vector Y_0 can be expanded in terms of this incomplete system (5.17)

$$Y_0 = \sum_{i=1}^{n} \alpha_i^0 \psi_i,$$

we have the inequality

$$\gamma_j = \bar{\xi}^{\mathrm{T}} \psi_j.$$

Substituting the value of γ_j into (5.26), we obtain

$$S = \sum_{j=n+1}^{l} \gamma_j^2 = \sum_{j=n+1}^{l} \left(\sum_{i=1}^{l} \xi_i \psi_j(x_i) \right)^2 = \sum_{j=n+1}^{l} \xi_j^2,$$

and hence the statistic S is distributed according to the central $\sigma^2 \chi^2$ distribution with $l - n$ degrees of freedom.

§4 A Theorem on Estimating the Mean Vector of a Multivariate Normal Distribution

In this section we shall obtain a family of estimators of the mean vector which are uniformly better than the estimator $\alpha(x, S) = x$. The estimator (5.21) belongs to this class.

Let x be a random vector distributed according to $N(\alpha_0, \sigma^2 I)$, and S be a random variable independent of x distributed according to the central $\sigma^2 \chi^2$ distribution with q degrees of freedom. We denote $F = x^{\mathrm{T}} x / S$.

The following theorem is valid.

Theorem 5.4 (Baranchik). *An estimator of the n-dimensional* ($n \geq 3$) *mean vector*

$$\hat{\alpha}(x, S) = \left(1 - \frac{r(F)}{F}\right) x,$$

where $r(F)$ is a monotonic nondecreasing function satisfying

$$0 \le r(F) \le 2\frac{n-2}{q+2}, \tag{5.27}$$

is uniformly better than the estimator $\alpha(x, S) = x$.

Remark. Theorem 5.2 is a particular case of Theorem 5.4 obtained by setting

$$r(F) = \begin{cases} \dfrac{n-2}{q+2} & \text{for } F \ge \dfrac{n-2}{q+2}, \\ F & \text{for } F < \dfrac{n-2}{q+2}. \end{cases}$$

PROOF. In the proof of Theorem 5.4 the following fact is used: the mathematical expectation of a random variable $f(\chi^2(n, b))$ taken with respect to the measure $\mu(\chi^2(n, b))$, where $\chi^2(n, b)$ is a random variable with the noncentral χ^2 distribution with n degrees of freedom and noncentrality parameter b, can be represented as

$$Mf(\chi^2(n, b)) = Mf(\chi^2_{n+2k}),$$

where χ^2_{n+2k} is a random variable with the central χ^2 distribution with $n + 2k$ degrees of freedom, and k is a random variable distributed according to the Poisson distribution with parameter b:

$$P(k) = \exp\{-b\}\frac{b^k}{k!}.$$

(The mathematical expectation on the right-hand side is evaluated with respect to x as well as with respect to k.)

Thus

$$Mf(\chi^2(n, b)) = Mf(\chi^2_{n+2k}) = \exp\{-b\}\sum_{t=0}^{\infty}\frac{b^t}{t!}Mf(\chi^2_{n+2t}). \tag{5.28}$$

We now proceed directly to the proof of the theorem: Our purpose is to show that the difference

$$H = M\|\hat{\alpha}(x, S) - \alpha_0\|^2 - M\|x - \alpha_0\|^2 \tag{5.29}$$

is nonnegative. Denote

$$g(F) = 1 - \frac{r(F)}{F}$$

and transform (5.29)

$$H = M[x^{\mathrm{T}}xg^2(F)] - 2\alpha_0^{\mathrm{T}}Mg(F)x + \|\alpha_0\|^2 - n\sigma^2. \tag{5.30}$$

The expressions (5.31)–(5.34) below are derived under the assumption that S is fixed. According to (5.28) we have

$$M\left[x^{\mathrm{T}}xg^2\left(\frac{x^{\mathrm{T}}x}{S}\right)\right] = \exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\sum_{t=0}^{\infty}\frac{\|\alpha_0\|^{2t}}{t!(2\sigma^2)^t}M\left[\sigma^2\chi^2_{n+2t}\,g^2\left(\frac{\sigma^2\chi^2_{n+2t}}{S}\right)\right]. \quad (5.31)$$

We now transform the expression

$$\alpha_0^{\mathrm{T}}Mg(F)x = \alpha_0^{\mathrm{T}}Mg\left(\frac{x^{\mathrm{T}}x}{S}\right)x.$$

For this purpose we shall perform an orthogonal transformation of vectors x into vectors z such that in the new coordinate system the mean vector is equal to $(\|\alpha_0\|, 0, \ldots, 0)$ (only the first coordinate does not vanish, and it is equal to the norm of the mean vector). This transformation leaves S unaltered. We obtain

$$\alpha_0^{\mathrm{T}}Mg\left(\frac{x^{\mathrm{T}}x}{S}\right)x = \|\alpha_0\|Mg\left(\frac{z^{\mathrm{T}}z}{S}\right)z_1,$$

where z is the first coordinate of the vector $z = (z_1, \ldots, z_n)^{\mathrm{T}}$.

Observe now that

$$M\left[g\left(\frac{z^{\mathrm{T}}z}{S}\right)z_1\right] = \frac{\sigma^2}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\} \times \frac{d}{d\|\alpha_0\|}\int g\left(\frac{\sum_{i=1}^{n}z^2}{S}\right)\exp\left\{-\frac{\sum_{i=1}^{n}z_i^2 - 2\|\alpha_0\|z_1}{2\sigma^2}\right\}dz_1\cdots dz_n.$$

Thus we obtain

$$\|\alpha_0\|Mg\left(\frac{z^{\mathrm{T}}z}{S}\right)z_1 = \frac{\sigma^2\|\alpha_0\|}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\} \times \frac{d}{d\|\alpha_0\|}\int g\left(\frac{\sum_{i=1}^{n}z_i^2}{S}\right)\exp\left\{-\frac{\sum_{i=1}^{n}z_i^2 - 2\|\alpha_0\|z_1}{2\sigma^2}\right\}dz_1\cdots dz_n$$

$$= \sigma^2\|\alpha_0\|\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\frac{d}{d\|\alpha_0\|}\exp\left\{\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}Mg\left(\frac{\sigma^2\chi^2_{n+2k}}{S}\right),$$

where k is a random variable distributed according to the Poisson distribution with the mean $\|\alpha_0\|^2/(2\sigma^2)$. Finally we obtain

$$\alpha_0^{\mathrm{T}}Mg\left(\frac{x^{\mathrm{T}}x}{S}\right)x = 2\sigma^2\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\sum_{t=0}^{\infty}t\left(\frac{\|\alpha_0\|^2}{2\sigma^2}\right)^t\frac{Mg(\sigma^2\chi^2_{n+2t}|S)}{t!}. \quad (5.32)$$

Now taking into account that $\|\alpha_0\|^2/(2\sigma^2)$ is the mean of the random variable k distributed according to the Poisson distribution, we express the third summand in the sum (5.30) in the form

$$\|\alpha_0\|^2 = 2\sigma^2 \exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\} \sum_{t=0}^{\infty} t \frac{\left(\frac{\|\alpha_0\|^2}{2\sigma^2}\right)^t}{t!}. \tag{5.33}$$

We can thus represent the expression (5.30) in the form

$$H = \sigma^2 \exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\} \sum_{t=0}^{\infty} \frac{\left(\frac{\|\alpha_0\|^2}{2\sigma^2}\right)^t}{t!}$$
$$\times \left[M\chi^2_{n+2t}\, g^2\left(\frac{\sigma^2\chi^2_{n+2t}}{S}\right) - 4tMg\left(\frac{\sigma^2\chi^2_{n+2t}}{S}\right) - n + 2t\right]. \tag{5.34}$$

Now let $S = \sigma^2\chi^2_q$ be a random variable distributed according to the central $\sigma^2\chi^2$ distribution with q degrees of freedom. The theorem will be proved if we verify that the expression

$$h = M\left[\chi^2_{n+2t} g^2\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right) - 4tg\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right) - n + 2t\right] \tag{5.35}$$

is nonpositive for all t.

Denote $\chi^2_{n+2t}/\chi^2_q = u$, and observe that

$$u(1 - g(u)) = r(u). \tag{5.36}$$

Therefore condition (5.27) implies that

$$g(u) > 1 - 2\frac{n-2}{q+2}u^{-1}. \tag{5.37}$$

We transform the expression (5.35) utilizing notation (5.36) and the fact that $M\chi^2_{n+2t} = n + 2t$:

$$h = M\left[-2r(u)\chi^2_q + r(u)(1 - g(u))\chi^2_q + 4t\frac{r(u)}{u}\right]$$
$$= M\left[r(u)\chi^2_q\left(-1 - g(u) + \frac{4t}{\chi^2_{n+2t}}\right)\right].$$

Taking (5.37) into account, we obtain that the quantity h does not exceed

$$\hat{h} = M(r(u)\zeta) = M\left[M\left\{r\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right)\zeta \,\middle|\, \chi^2_q\right\}\right],$$

where

$$\zeta = \chi^2_q\left[-2 + \left(4t + 2\frac{n-2}{q+2}\chi^2_q\right)\frac{1}{\chi^2_{n+2t}}\right].$$

For any fixed χ_q^2 we determine a constant a such that

$$-2 + \left(4t + 2\frac{n-2}{q+2}\chi_q^2\right)\frac{1}{a} = 0. \tag{5.38}$$

Observe that for any $\chi_{n+2t}^2 > a$ the inequality $\zeta < 0$ is valid. Therefore taking into account that in view of the condition of the theorem the function $r(u)$ is nondecreasing, we obtain the bound

$$\begin{aligned}
&M\left\{r\left(\frac{\chi_{n+2t}^2}{\chi_q^2}\right)\zeta \,\middle|\, \chi_q^2\right\} \\
&\quad \le r\left(\frac{a}{\chi_q^2}\right)M\{\zeta \mid \chi_{n+2t}^2 \le a\}P\{\chi_{n+2t}^2 \le a\} \\
&\qquad + r\left(\frac{a}{\chi_q^2}\right)M\{\zeta \mid \chi_{n+2t}^2 > a\}P\{\chi_{n+2t}^2 > a\} \\
&\quad = r\left(\frac{a}{\chi_q^2}\right)M\{\zeta \mid \chi_q^2\} \\
&\quad = r\left(\frac{a}{\chi_q^2}\right)\chi_q^2\left[-2 + \left(4t + 2\frac{n-2}{q+2}\chi_q^2\right)\frac{1}{n+2t-2}\right] \\
&\quad = 2\frac{n-2}{n+2t-2}r\left(\frac{a}{\chi_q^2}\right)\chi_q^2\left(-1 + \frac{\chi_q^2}{q+2}\right).
\end{aligned} \tag{5.39}$$

(We have used the equality $M(1/\chi_m^2) = 1/(m-2)$ $(m \ge 3)$.)

Substitute now into (5.39) the value of a satisfying (5.38), and compute the mathematical expectation of the last term in (5.39), which is

$$2\frac{n-2}{n+2t-2}M\left\{r\left(\frac{2t}{\chi_q^2} + \frac{n-2}{q+2}\right)\chi_q^2\left[-1 + \frac{\chi_q^2}{q+2}\right]\right\}.$$

Taking into account that $r(u)$ is a nondecreasing function we find the bound

$$\begin{aligned}
&M\left\{r\left(\frac{2t}{\chi_q^2} + \frac{n-2}{q+2}\right)\chi_q^2\left[-1 + \frac{\chi_q^2}{q+2}\right]\right\} \\
&\quad \le r\left(\frac{n+2t-2}{q+2}\right)M\left\{\chi_q^2\left[-1 + \frac{\chi_q^2}{q+2}\right] \,\middle|\, \chi_q^2 \le q+2\right\} \\
&\qquad + r\left(\frac{n+2t-2}{q+2}\right)M\left\{\chi_q^2\left(-1 + \frac{\chi_q^2}{q+2}\right) \,\middle|\, \chi_q^2 > q+2\right\} \\
&\quad = r\left(\frac{n+2t-2}{q+2}\right)M\left\{\chi_q^2\left[-1 + \frac{\chi_q^2}{q+2}\right]\right\} = 0.
\end{aligned}$$

(For a central χ^2 distribution we have $M\chi_q^2 = q$, $M(\chi_q^2)^2 = q(q+2)$.)

Thus the quantity (5.35) is nonpositive and the theorem is proved. □

§5 The Gauss–Markov Theorem

Up until now, when estimating regression it was assumed that the errors are distributed according to the normal distribution. We shall now relax this assumption. It will be assumed that the distribution of errors is unknown but has a bounded variance. Under these conditions it is required to construct the best algorithm for the regression estimation.

Above, when developing the theory of normal regression we first established that in the class of algorithms leading to unbiased estimators of the parameters the least-squares method was optimal, but for a wider class of algorithms procedures which are better than the least-squares method were obtained. We shall now proceed analogously. First we shall show that in some narrow class of estimating algorithms the least-squares method is the best, and then we obtain estimation methods in a wider class of algorithms which are superior to the least-squares method.

Under the assumption of normal errors the least-squares method is the best in the class of unbiased methods of estimation. In this section we shall show that in a narrower class of estimates which are both linear and unbiased, the least-squares method yields the best estimating algorithms independently of the distribution of the errors.

Definition. We say that an estimator of the parameter α is *linear* in the observations $Y = (y_1, \ldots, y_l)^T$ if it can be represented in the form

$$\alpha = LY \qquad \left(\alpha_j = \sum_{i=1}^{l} \beta_{ij} y_i\right), \tag{5.40}$$

where L is a matrix with the entries β_{ij} $(i = 1, \ldots, l; j = 1, \ldots, n)$.

The following theorem is valid:

Theorem (Gauss–Markov). *Among all the linear unbiased estimators the least-squares estimator possesses the minimal variances of the coordinates.*

We shall prove the Gauss–Markov theorem in its more general form for the case of linear biased estimators. Denote by α_0 the vector of parameters of the linear regression

$$MY = \Phi\alpha_0 \qquad (Y = \Phi\alpha_0 + \bar{\xi}). \tag{5.41}$$

Define the estimator $\alpha(B)$ as the solution of the equation

$$(\Phi^T\Phi + B)\alpha(B) = \Phi^T Y, \tag{5.42}$$

where B is a symmetric nonnegative definite $n \times n$ matrix which defines the bias vector μ_0. We shall show that the estimator $\alpha(B)$ possesses extremal properties. Namely, the following theorem is valid.

Theorem 5.5. *Among all the linear estimators of the vector of parameters α with the bias vector equaling μ_0, the estimator $\alpha(B)$ possesses the minimal variance of coordinates.*

PROOF. We obtain from (5.42)

$$M\alpha(B) = M(\Phi^T\Phi + B)^{-1}\Phi^T Y = (\Phi^T\Phi + B)^{-1}\Phi^T\Phi\alpha_0. \tag{5.43}$$

Let $\hat{\alpha} = LY$ be an arbitrary linear estimator such that

$$M\hat{\alpha} = M\alpha(B) = \mu_0 + \alpha_0 = \mu. \tag{5.44}$$

Then we obtain from (5.42)

$$MLY = L\Phi\alpha_0 = (\Phi^T\Phi + B)^{-1}\Phi^T\Phi\alpha_0. \tag{5.45}$$

Since the equality (5.45) is valid for any α_0, then

$$L\Phi = (\Phi^T\Phi + B)^{-1}\Phi^T\Phi. \tag{5.46}$$

We now write the variance of the ith coordinate of estimator $\hat{\alpha}$:

$$\begin{aligned} M(\hat{\alpha}_i - \mu_i)^2 &= M(\hat{\alpha}_i - \alpha_i(B) + \alpha_i(B) - \mu_i)^2 \\ &\geq M(\alpha_i(B) - \mu_i)^2 + 2M(\hat{\alpha}_i - \alpha_i(B))(\alpha_i(B) - \mu_i), \end{aligned} \tag{5.47}$$

where μ_i is the ith coordinate of the vector μ.

We shall show that the second summand on the right-hand side of (5.47) vanishes. Indeed, utilizing (5.44) and (5.46), we obtain

$$\begin{aligned} &M(\hat{\alpha}_i - \alpha_i(B))(\alpha_i(B) - \mu_i) \\ &\quad = M(\hat{\alpha}_i - \alpha_i(B))\alpha_i(B) \\ &\quad = \sigma^2\|(L - (\Phi^T\Phi + B)^{-1}\Phi^T)\Phi(\Phi^T\Phi + B)^{-1}\|_{ii} \\ &\quad = \sigma^2\|(L\Phi - (\Phi^T\Phi + B)^{-1}\Phi^T\Phi)(\Phi^T\Phi + B)^{-1}\|_{ii} = 0, \end{aligned}$$

where $\|A\|_{ii}$ denotes the element A_{ii} of the matrix $\|A\|$.

Thus

$$M(\hat{\alpha}_i - \mu_i)^2 \geq M(\alpha_i(B) - \mu_i)^2.$$

The theorem is thus proved. □

The Gauss–Markov theorem follows from the theorem just proved by setting $\|B\| = \|0\|$ in (5.42). In that case $\mu_0 = 0$.

Further, in Chapter 8 to construct the regression estimators from small samples we shall make use of this theorem. We shall search for the best estimators among the estimators of the class $\alpha(\gamma B)$ (where $\gamma > 0$ is a constant specifying the estimator of the class). The estimator $\alpha(\gamma^* B)$ is called a *ridge-regression estimator.*

§6 Best Linear Estimators

Thus, among linear unbiased estimators, the least-squares estimators are the best regardless of the distribution of the errors. In the next sections we shall consider a wider class of estimators—the class of linear estimators (not necessarily unbiased), and we shall obtain the best estimators in this class. These estimators will differ from the least-squares estimators provided nontrivial prior information concerning the estimated parameters is available. In cases when no nontrivial prior information is available the best linear estimator is still the least-squares method.

Let the parameters of the regression

$$\bar{y} = y(x) = \sum_{i=1}^{n} \alpha_i^0 \psi_i(x) \tag{5.48}$$

in a Gauss–Markov model be estimated from empirical data $x_1, y_1, \ldots, x_l, y_l$. Let $\hat{\psi}_1(x), \ldots, \hat{\psi}_n(x)$ be a doubly orthogonal basis

$$\begin{aligned} \int \hat{\psi}_i(x)\hat{\psi}_j(x)P(x)\,dx &= \begin{cases} \lambda_i & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \\ \sum_{r=1}^{l} \hat{\psi}_i(x_r)\hat{\psi}_j(x_r) &= \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases} \end{aligned} \tag{5.49}$$

Consider the class of linear estimators:

$$\hat{\alpha}_p = \theta_p^{\mathrm{T}} Y + \beta_0^p, \tag{5.50}$$

where

$$\theta_p = (\theta_1^p, \ldots, \theta_l^p)^{\mathrm{T}}, \qquad Y = (y_1, \ldots, y_l)^{\mathrm{T}}.$$

We introduce the system of orthogonal vectors

$$\chi_1, \ldots, \chi_l; \qquad \chi_i^{\mathrm{T}} \chi_j = \begin{cases} l & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \tag{5.51}$$

in which the first n vectors are

$$\chi_i = (\hat{\psi}_i(x_1), \ldots, \hat{\psi}_i(x_l))^{\mathrm{T}}, \qquad i = 1, \ldots, n.$$

We represent the vector θ_p in the expansion in terms of (5.51):

$$\theta_p = \sum_{i=1}^{l} \beta_i^p \chi_i. \tag{5.52}$$

Then the equality (5.50) can be rewritten as

$$\hat{\alpha}_p = \sum_{i=1}^{l} \beta_i^p \chi_i^{\mathrm{T}} Y + \beta_0^p. \tag{5.53}$$

We express the amount of deviation $M(\hat{\alpha}_p - \alpha_p^0)^2$ in terms of the parameters β. For this purpose we shall utilize the identity

$$M(\hat{\alpha}_p - \alpha_p^0)^2 = (M(\hat{\alpha}_p - \alpha_p^0))^2 + M(\hat{\alpha}_p - M\hat{\alpha}_p)^2. \tag{5.54}$$

The first summand on the right-hand side equals

$$(M(\hat{\alpha}_p - \alpha_p^0))^2 = \left(l\sum_{i=1}^{n} \beta_i^p \alpha_i^0 + \beta_0^p - \alpha_p^0\right)^2.$$

The second summand equals

$$M(\hat{\alpha}_p - M\hat{\alpha}_p)^2 = l\sigma^2 \sum_{i=1}^{l} (\beta_i^p)^2.$$

Thus

$$M(\hat{\alpha}_p - \alpha_p^0)^2 = \sigma^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \left(l\sum_{i=1}^{n} \beta_i^p \alpha_i^0 + \beta_0^p - \alpha_p^0\right)^2 = \mathscr{D}^p(\beta \mid \alpha, \sigma). \tag{5.55}$$

The best linear estimator is the estimator which minimizes (5.55).

§7 Criteria for the Quality of Estimators

The best linear estimator can be obtained by directly minimizing with respect to $\beta_1, \ldots, \beta_l$ the right-hand side of the equality (5.55). The minimum of (5.55) is attained at $\beta_1^p = \beta_2^p = \cdots = \beta_l^p = 0$ and $\beta_0^p = \alpha_p^0$, and this minimum is zero.

Thus for each specific problem (specific values of α_0 and σ) a trivial biased estimator can be found which yields the minimum of the square of deviations. Now we wish to construct a linear estimator which will be suitable for a solution of a class of problems rather than for a single one.

Let us define a class of problems $R(a, \sigma)$, to which the algorithm is applicable, by means of the inequalities

$$\begin{aligned} a_p \le \alpha_p \le b_p, \\ d \le \sigma \le e. \end{aligned} \tag{5.56}$$

We shall now determine the quality of an algorithm for estimating the parameter α_p in the class $R(\alpha, \sigma)$. As usual in such situations, we shall consider two approaches: Bayesian and minimax. For each approach a different notion of the quality of a linear estimator will be introduced.

According to Bayes's principle the best method for estimation is that for which the mean value of the criterion over the set of problems belonging to $R(\alpha, \sigma)$ is minimal (the measure on this set is given by the distribution $P(\alpha, \sigma)$).

Definition. The estimator

$$\alpha_p^{(1)} = \sum \beta_i^p \chi_i^T Y + \beta_0^p$$

is called *linearly best in the mean* if among all linear estimators it yields the minimum of the functional

$$\mathscr{D}_1^p(\beta) = \int \mathscr{D}^p(\beta|\alpha, \sigma) P(\alpha, \sigma)\, d\alpha_1 \cdots d\alpha_n\, d\sigma. \tag{5.57}$$

Below we shall compute a Bayesian estimator for the case when the parameters α and σ are distributed independently according to the uniform distribution on the corresponding intervals, i.e.,

$$P(\alpha, \sigma) = \begin{cases} \prod_{i=1}^{n} \frac{1}{b_i - a_i} \frac{1}{e - d} & \text{if } a_p \le \alpha_p \le b_p,\ d \le \sigma \le e, \\ 0 & \text{otherwise.} \end{cases} \tag{5.58}$$

Thus the quality of the estimator is determined by the functional

$$\mathscr{D}_1^p(\beta) = \int \mathscr{D}^p(\beta|\alpha, \sigma) \prod_{i=1}^{n} \frac{d\alpha_i}{b_i - a_i} \frac{d\sigma}{e - d}. \tag{5.59}$$

In accordance with the minimax principle the best method of estimation is considered to be the one which yields the minimum of $\mathscr{D}^p(\beta|\alpha, \sigma)$ for the least favorable problem (pair α, σ).

Definition. The estimator

$$\alpha_p^{(2)} = \sum \beta_i^p \chi_i^T Y + \beta_0^p$$

is called the *best linear minimax estimator* in the class $R(\alpha, \sigma)$ if it yields the minimum of the functional

$$\mathscr{D}_2^p(\beta) = \sup_{\alpha, \sigma} \mathscr{D}^p(\beta|\alpha, \sigma) \tag{5.60}$$

in the class of linear estimators.

In general there may exist problems belonging to the class $R(\alpha, \sigma)$ for which the estimators $\alpha_p^{(1)}$ and $\alpha_p^{(2)}$ introduced above are worse than the least-squares estimators $\beta_{\text{lse}}^p = (0, \ldots, 1/l, \ldots, 0)^T$, $\beta_0^p = 0$ (only the pth coordinate of the vector β_{lse}^p is nonvanishing). Therefore we shall define the third optimal estimator in such a manner that it will be uniformly better than the least-squares estimator. For this purpose we introduce the loss function

$$\mathscr{D}_3^p(\beta) = \sup_{\alpha, \sigma} (\mathscr{D}^p(\beta|\alpha, \sigma) - \mathscr{D}^p(\beta_{\text{lse}}|, \alpha, \sigma)) \tag{5.61}$$

and require that the optimal estimator minimize the expression (5.61).

Definition. The estimator

$$\alpha_p^{(3)} = \sum \beta_i^p \chi_i^T Y + \beta_0^p$$

is called *linearly uniformly better* than the least-squares estimator if it yields the minimum of the functional (5.61) in the class of linear estimators and $\min_\beta D_3^p(\beta) < 0$.

§8 Evaluation of the Best Linear Estimators

The following three theorems constitute the basic content of the theory of the best linear estimator.

Theorem 5.6 (Koshcheev). *The best linear estimator of parameter α_p in the class $R(\alpha, \sigma)$ is of the form*

$$\alpha_p^{(i)} = \frac{\alpha_{\text{lse}}^p + \frac{c_p}{l}\rho_p^{(i)}}{1 + \frac{1}{l}\rho_p^{(i)}}, \qquad i = 1, 2, 3, \tag{5.62}$$

where $c_p = (a_p + b_p)/2$, α_{lse}^p is the least-squares estimator, $\alpha_p^{(1)}$ is the best in the mean estimator,

$$\rho_p^{(1)} = 4\frac{d^2 + de + e^2}{(a_p - b_p)^2}, \tag{5.63}$$

$\alpha_p^{(2)}$ is the best minimax estimator,

$$\rho_p^{(2)} = 4\frac{e^2}{(a_p - b_p)^2}, \tag{5.64}$$

$\alpha_p^{(3)}$ is the uniformly best estimator, and

$$\rho_p^{(3)} = 4\frac{d^2}{(a_p - b_p)^2}. \tag{5.65}$$

It thus turns out that the best linear estimators are biased. The structure of the estimators is given by the expression (5.62), where $\rho_p^{(i)}$ are defined in (5.63)–(5.65), depending on the specific notion of the quality of an estimator. There exists a simple relationship which shows by how much a Bayes or minimax estimator is superior to a least-squares estimator.

Theorem 5.7 (Koshcheev). *The equality*

$$\mathscr{D}_i^p(\alpha_p^{(i)}) = \frac{1}{1 + \frac{1}{l}\rho_p^{(i)}}\,\mathscr{D}_i^p(\alpha_{\text{lse}}^p), \qquad i = 1, 2 \tag{5.66}$$

is valid.

According to Theorem 5.7 the optimal estimators $\alpha_p^{(i)}$ are superior to the least-squares estimator by the factor $(1 + (1/l)\rho_p^{(i)})$. Hence the smaller the sample size l, the better the estimators $\alpha_p^{(i)}$.

Below we shall present the proof of Theorem 5.6. The validity of Theorem 5.7 follows from a more general theorem considered in the next section.

PROOF OF THEOREM 5.6

(1) *Derivation of the best linear estimator in the mean.* We write the functional whose minimum determines under our conditions the best estimator in the mean:

$$\mathscr{D}_1^p(\beta) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \int_d^e \left[l\sigma^2 \sum_{i=1}^{l} (\beta_i^p)^2 + \left(l \sum_{i=1}^{n} \beta_i^p \alpha_i + \beta_0^p - \alpha_p \right)^2 \right] \prod_{i=1}^{n} \frac{d\alpha_i}{b_i - a_i} \frac{d\sigma}{e - d}. \tag{5.67}$$

This integral can be easily evaluated:

$$\mathscr{D}_1^p(\beta) = \frac{l}{3} \frac{e^3 - d^3}{e - d} \sum_{i=1}^{l} (\beta_i^p)^2 + \prod_{j=1}^{n} \frac{1}{(b_j - a_j)} \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \left(l \sum_{i=1}^{n} \beta_i^p \alpha_i + \beta_0^p - \alpha_p \right)^2 d\alpha_1 \cdots d\alpha_n.$$

Denoting $(a_i + b_i)/2 = c_i, (a_i - b_i)/2 = \mathscr{M}_i, t_i = \alpha_i - c_i$, and substituting the variables, we obtain

$$\mathscr{D}_1^p(\beta) = \frac{l}{3} \frac{e^3 - d^3}{e - d} \sum_{i=1}^{l} (\beta_i^p)^2 + \prod_{j=1}^{n} \frac{1}{2\mathscr{M}_j} \int_{-\mathscr{M}_1}^{\mathscr{M}_1} \cdots \int_{-\mathscr{M}_n}^{\mathscr{M}_n} \left(l \sum_{i=1}^{n} \beta_i^p (t_i + c_i) + \beta_0^p - (t_p + c_p) \right)^2 dt_1 \cdots dt_n. \tag{5.68}$$

Since the integration is carried out over the symmetric intervals $[-\mathscr{M}, \mathscr{M}]$, the terms linear in t vanish. We thus obtain

$$\mathscr{D}_1^p(\beta) = \frac{l}{3}(e^2 + ed + d^2) \sum_{i=1}^{l} (\beta_i^p)^2 + \left(\beta_0^p + \sum_{i=1}^{n} (l\beta_i^p - \delta_{ip}) c_i \right)^2 + \prod_{j=1}^{n} \frac{1}{2\mathscr{M}_j} \int_{-\mathscr{M}_1}^{\mathscr{M}_1} \cdots \int_{-\mathscr{M}_n}^{\mathscr{M}_n} \sum_{i=1}^{n} (l\beta_i^p - \delta_{ip})^2 t_i^2 \, dt_1 \cdots dt_n. \tag{5.69}$$

Here the notation

$$\delta_{ip} = \begin{cases} 1 & \text{for } i = p, \\ 0 & \text{for } i \neq p \end{cases}$$

is utilized. Finally we arrive at

$$\mathscr{D}_1^p(\beta) = \frac{l}{3}(e^2 + ed + d^2) \sum_{i=1}^{l} (\beta_i^p)^2 + \left(\beta_0^p + \sum_{i=1}^{n} (l\beta_i^p - \delta_{ip}) c_i \right)^2 + \sum_{i=1}^{n} \frac{\mathscr{M}_i^2}{3} (l\beta_i^p - \delta_{ip})^2. \tag{5.70}$$

In order to obtain the best linear estimator in the mean it remains only to minimize the expression (5.70) with respect to parameters β.

Equating the partial derivatives of (5.70) to zero, we obtain that

$$\beta_i^p = 0 \quad \text{for } i \neq p,$$

$$\beta_0^p = -c_p(l\beta_p^p - 1),$$

$$\beta_p^p = \frac{\dfrac{\mathscr{M}_p^2}{e^2 + ed + d^2}}{1 + \dfrac{l\mathscr{M}_p^2}{e^2 + ed + d^2}}. \tag{5.71}$$

Substituting the values (5.71) obtained into (5.53), we have

$$\alpha_p^{(1)} = \frac{\dfrac{\mathscr{M}_p^2}{e^2 + ed + d^2}}{1 + \dfrac{l\mathscr{M}_p^2}{e^2 + ed + d^2}} \chi_p^{\mathrm{T}} Y + \frac{c_p}{1 + \dfrac{l\mathscr{M}_p^2}{e^2 + ed + d^2}}.$$

Introduce the notation $\rho_p^{(1)} = (e^2 + ed + d^2)/\mathscr{M}_p^2$. Then

$$\alpha_p^{(1)} = \frac{\dfrac{1}{l}\chi_p^{\mathrm{T}} Y + \dfrac{c_p}{l}\rho_p^{(1)}}{1 + \dfrac{1}{l}\rho_p^{(1)}}.$$

Observe that the quantity $(1/l)\chi_p^{\mathrm{T}} Y$ is the least-squares estimator of the parameter α_p^0. Thus

$$\alpha_p^{(1)} = \frac{\alpha_{\mathrm{se}}^p + \dfrac{c_p}{l}\rho_p^{(1)}}{1 + \dfrac{1}{l}\rho_p^{(1)}}.$$

The first part of the theorem is proved.

(2) *Derivation of the best minimax estimator.* The functional whose minimum determines the best minimax estimator is equal to

$$\mathscr{D}_2^p(\beta) = \sup_{\sigma,\alpha}\left[\sigma^2 l\sum_{i=1}^{l}(\beta_i^p)^2 + \left(l\sum_{i=1}^{n}(\beta_i^p\alpha_i + \beta_0^p - \alpha_p)\right)^2\right]. \tag{5.72}$$

Utilizing the notation

$$c_i = \frac{b_i + a_i}{2}, \qquad \mathscr{M}_i = \frac{b_i - a_i}{2}, \qquad t_i = \alpha_i - c_i,$$

and substituting the variables in (5.72), we have

$$\begin{aligned}
\mathscr{D}_2^p(\beta) &= e^2 l\sum_{i=1}^{l}(\beta_i^p)^2 + \sup_{|t_i|\le\mathscr{M}_i}\left[\sum_{i=1}^{n}(l\beta_i^p - \delta_{ip})(t_i + c_i) + \beta_0^p\right]^2 \\
&= e^2 l\sum_{i=1}^{l}(\beta_i^p)^2 + \sup_{|t_i|\le\mathscr{M}_i}\left[\sum_{i=1}^{n}(l\beta_i^p - \delta_{ip})t_i + \sum_{i=1}^{n}(l\beta_i^p - \delta_{ip})c_i + \beta_0^p\right]^2 \\
&= e^2 l\sum_{i=1}^{l}(\beta_i^p)^2 + \left[\sum_{i=1}^{n}|l\beta_i^p - \delta_{ip}|\mathscr{M}_i + \left|\sum_{i=1}^{n}(l\beta_i^p - \delta_{ip})c_i + \beta_0^p\right|\right]^2.
\end{aligned}$$

Thus

$$\mathscr{D}_2^p(\beta) = e^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \left[\sum_{i=1}^{n} |l\beta_i^p - \delta_{ip}| \mathscr{M}_i + \left| \sum_{i=1}^{n} (l\beta_i^p - \delta_{ip}) c_i + \beta_0^p \right| \right]^2. \quad (5.73)$$

We shall now obtain the minimum of (5.73). By choosing β_0^p to be equal to

$$\beta_0^p = -\sum_{i=1}^{n} (l\beta_i^p - \delta_{ip}) c_i,$$

the second term of the sum in the square brackets becomes zero. Therefore it is sufficient to minimize

$$\mathscr{D}_2^p(\beta) = e^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \left(\sum_{i=1}^{n} |l\beta_i^p - \delta_{ip}| \mathscr{M}_i \right)^2. \quad (5.74)$$

The minimum of (5.74) is attained for

$$\beta_i^p = 0 \quad \text{for } i \neq p, \quad (5.75)$$

whence for $\beta_i^p = 0$ $(i \neq p)$ the functional (5.74) becomes

$$\mathscr{D}_2^p(\beta)|_{\beta_i^p = 0 \, (i \neq p)} = l e^2 (\beta_p^p)^2 + (l\beta_p^p - 1)^2 \mathscr{M}_p^2. \quad (5.76)$$

The minimum of this expression is attained at

$$\beta_p^p = \frac{\mathscr{M}_p^2}{e^2 + l\mathscr{M}_p^2}. \quad (5.77)$$

Substituting (5.75) and (5.77) into (5.53), we obtain the best minimax estimator

$$\alpha_p^{(2)} = \frac{\mathscr{M}_p^2}{e^2 + l\mathscr{M}_p^2} \chi_p^{\mathrm{T}} Y + \left(\frac{l\mathscr{M}_p^2}{e^2 + l\mathscr{M}_p^2} - 1 \right) c_p = \frac{l\mathscr{M}_p^2 \alpha_{\mathrm{lse}}^p + c_p e^2}{e^2 + l\mathscr{M}_p^2}.$$

Introducing the notation $\rho_p^{(2)} = e^2 / l\mathscr{M}_p^2$, we arrive at

$$\alpha_p^{(2)} = \frac{\alpha_{\mathrm{lse}}^p + \frac{c_p}{l} \rho_p^{(2)}}{1 + \frac{1}{l} \rho_p^{(2)}}.$$

(3) *Derivation of the uniformly best linear estimator.* To evaluate the uniformly best estimator it is required to minimize the functional

$$\mathscr{D}_3^p(\beta) = \sup_{\alpha, \sigma} (\mathscr{D}^p(\beta | \alpha, \sigma) - \mathscr{D}^p(\beta_{\mathrm{lse}} | \alpha, \sigma)),$$

or explicitly,

$$\mathscr{D}_3^p(\beta) = \sup_{d \leq \sigma \leq e} \left[l\sigma^2 \left(\sum_{i=1}^{l} (\beta_i^p)^2 - 1 \right) \right] + \sup_{a_i \leq \alpha_i \leq b_i} \left[\sum_{i=1}^{n} (l\beta_i^p - \delta_{ip}) \alpha_i + \beta_0^p \right]^2. \quad (5.78)$$

It is easy to verify that in this case all the calculations are the same as those carried out in the preceding subsection, except that if

$$\sum_{i=1}^{l} (\beta_i^p)^2 - 1 < 0, \tag{5.79}$$

then $d = \inf \sigma$ should be taken instead of $e = \sup \sigma$.

Consequently

$$\beta_0^p = -\sum_{i=1}^{l} (l\beta_i^p - \delta_{ip})c_i,$$

$$\beta_i^p = \begin{cases} 0 & \text{for } i \neq p, \\ \dfrac{\mathscr{M}_i^2}{s^2 + l\mathscr{M}_i^2} & \text{for } i = p, \end{cases} \tag{5.80}$$

where s is either $\inf \sigma$ or $\sup \sigma$, depending on the sign of $\sum_{i=1}^{n} (\beta_i^p)^2 - 1$. However, for β_i^p as given by (5.80) the expression (5.79) is negative:

$$\sum_{i=1}^{l} (\beta_i^p)^2 - 1 = \left(\frac{\mathscr{M}_p^2}{s^2 + \mathscr{M}_p^2 l} \right)^2 - 1 < 0.$$

Hence $s = \inf \sigma = d$. Thus the uniformly best linear estimator is equal to

$$\alpha_p^{(3)} = \frac{\alpha_{\text{lse}}^p + \dfrac{c_p}{l} \rho_p^{(3)}}{1 + \dfrac{1}{l} \rho_p^{(3)}},$$

where in this case

$$\rho_p^{(3)} = \frac{d^2}{\mathscr{M}_p^2}. \qquad \square$$

§9 Utilizing Prior Information

According to Theorem 5.6 the availability of the following prior information:

(1) the interval $[a_i, b_i]$ to which the estimated parameter α_p belongs,
(2) the interval $[d, e]$ to which the variance of the noise σ belongs,

allows us to construct the best linear estimators. According to Theorem 5.7 the functional defining the quality of the best linear estimator is $1 + (\rho_p^{(i)}/l)$ times smaller than the functional corresponding to the least-squares estimator.

Usually it is not too difficult to obtain this prior information for solving practical problems within the Gauss–Markov model. As a rule the intervals in which the measured values of y are situated,

$$\tau_i \leq y_i \leq T_i \tag{5.81}$$

are known. This knowledge results from long experience or from the knowledge of the laws of nature. For example, when constructing the regression for the temperature forecast in Moscow on the 166th day of the year, it is known *a priori* that the forecast value of t lies within the given limits $+5°C \leq t \leq 35°C$. The knowledge of the bounds (5.81) allows us to obtain intervals for the estimated parameters. The equality $\alpha_p^0 = M(1/l)\chi_p^T Y$ implies that

$$b_p = \sup_Y \frac{1}{l} \chi_p^T Y \leq \frac{1}{l} \left(\sum_{i=1}^{l}{}' T_i \hat{\psi}_p(x_i) + \sum_{i=1}^{l}{}'' \tau_i \hat{\psi}_p(x_i) \right).$$

Here the first sum $\sum'$ contains the positive coordinates of the vector $\chi_p = (\hat{\psi}_p(x_1), \ldots, \hat{\psi}_p(x_l))^T$, while the second contains the negative ones. Analogously the bounds

$$a_p = \inf_Y \frac{1}{l} \chi_p^T Y \geq \frac{1}{l} \left(\sum_{i=1}^{l}{}' \tau_i \hat{\psi}_p(x_i) + \sum_{i=1}^{l}{}'' T_i \hat{\psi}_p(x_i) \right)$$

are obtained.

To estimate the interval for the variance we can also utilize our experience and knowledge of the laws which govern errors. However, if the interval obtained for the variance is too wide, we can then use alternatively the probabilistic approach, which consists of choosing the interval which contains the true value of the variance with the highest probability.

It is known that the quantity

$$\sigma_{\text{emp}}^2 = \frac{\sum_{i=1}^{l} y_i^2 - l \sum_{p=1}^{n} (\alpha_{\text{lse}}^p)^2}{l - n}$$

is an unbiased estimator of the error variance. We shall utilize Chebyshev's inequality

$$P\left\{ \sigma_{\text{emp}}^2 \geq \frac{\sigma^2}{\eta} \right\} \leq \eta,$$

which implies that with probability $1 - \eta$

$$\sigma^2 \geq \sigma_{\text{emp}}^2 \eta. \tag{5.82}$$

The bound (5.82) may be refined if the nature of the error distribution is known.

Based on the interval for the variance $d \leq \sigma \leq e$ and the interval to which the parameter α_p belongs, the parameters $\rho_p^{(i)}$ and $c_p^{(i)}$ are found by means of which optimal linear estimators are constructed. Note that the more indefinite the prior information is (the wider the interval is), the smaller the value of $\rho_p^{(i)}$ will be and the closer the best linear estimator will be to the least-squares estimator. It can be shown that for trivial prior information ($-\infty < \alpha_p < \infty$, $0 < \sigma < \infty$) the best linear estimator coincides with the least-squares one.

To complete the theory of the best linear estimation it remains to clarify how sensitive the methods of linear estimation are to the precision of prior information. Theorem 5.8 answers this question.

Theorem 5.8 (Koshcheev). *Let* $\hat{\alpha}_p^i = \alpha_p(\hat{\rho}_p^{(i)}, \hat{c}_p)$ *be the best linear estimator computed from approximate values of the parameters* $\hat{\rho}_p^{(i)}, \hat{c}_p, \hat{\mathscr{M}}_p$, *while the true values of the parameters equal* $\rho_p^{(i)}, c_p, \mathscr{M}_p$. *Then the quality of the estimator obtained is given by*

$$\mathscr{D}_i^p(\hat{\alpha}_p(\hat{\rho}_p^{(i)}, \hat{c}_p)) = \frac{1 + v_i \dfrac{(\hat{\rho}_p^{(i)})^2}{\rho_p^{(i)}}}{(1 + \hat{\rho}_p^{(i)})^2} \mathscr{D}_i^p(\alpha_{\text{lse}}^p) \qquad (i = 1, 2), \tag{5.83}$$

where

$$v_1 = 1 + 3\left(\frac{\hat{c}_p - c_p}{\mathscr{M}_p}\right)^2, \qquad v_2 = \left(1 + \frac{|c_p - \hat{c}_p|}{\mathscr{M}_p}\right)^2. \tag{5.84}$$

Observe that Theorem 5.7 is a particular case of Theorem 5.8 for $\hat{c}_p = c_p$ and $\hat{\rho}_p^{(i)} = \rho_p^{(i)}$.

It follows from the equality (5.83) that if the value of parameter $\hat{\rho}_p^{(i)}$ is related to $\rho_p^{(i)}$ and v_i by the inequality

$$\rho_p^{(i)} > \frac{\hat{\rho}_p^{(i)} v_i}{2 + \hat{\rho}_p^{(i)}}, \tag{5.85}$$

then the estimator obtained using $\hat{\rho}_p^{(i)}, \hat{c}_p$ will be better than the least-squares estimator. Consequently the choice of $\hat{\rho}_p^{(i)}$ is based on two contradictory considerations. To obtain an estimator at least as good as the least-squares one, the value of $\hat{\rho}_p^{(i)}$ should be reduced (so that (5.85) is fulfilled). But the gain, which is approximately equal to $\mathscr{D}_i(\alpha_{\text{lse}}^p)/(1 + \hat{\rho}^{(i)})$, is decreased.

Proof of Theorem 5.8. First we shall compute the value of the criterion (5.55) for the estimator $\hat{\alpha}_p(\hat{\rho}_p^{(i)} c_p)$:

$$M(\alpha_p(\hat{\rho}_p^{(i)}, \hat{c}_p) - \alpha_p^0)^2$$

$$= M\left(\frac{\alpha_{\text{lse}}^p + \dfrac{\hat{c}_p}{l}\hat{\rho}_p{}^{(i)}}{1 + \dfrac{\hat{\rho}_p^{(i)}}{l}} - \frac{\alpha_p^0 + \dfrac{\hat{c}_p}{l}\hat{\rho}_p^{(i)}}{1 + \hat{\rho}_p^{(i)}\dfrac{1}{l}}\right)^2 + \left(\frac{\alpha_p^0 + \dfrac{\hat{c}_p}{l}\hat{\rho}_p^{(i)}}{1 + \hat{\rho}_p^{(i)}\dfrac{1}{l}} - \alpha_p^0\right)^2$$

$$= \frac{\dfrac{\sigma^2}{l}}{\left(1 + \dfrac{\hat{\rho}_p^{(i)}}{l}\right)^2} + \frac{\left(\dfrac{\hat{\rho}_p^{(i)}}{l}\right)^2 (\hat{c}_p - \alpha_p^0)^2}{\left(1 + \dfrac{\hat{\rho}_p^{(i)}}{l}\right)^2}.$$

The two relations (5.83) claimed in the theorem are verified by elementary calculations

$$\mathscr{D}_1^p(\hat{\alpha}) = \int_{c_p - \mathscr{M}_p}^{c_p + \mathscr{M}_p} \int_d^e \frac{\dfrac{\sigma^2}{l} + \left(\dfrac{\hat{\rho}_p^{(1)}}{l}\right)^2 (\hat{c}_p - \alpha_p^0)^2}{\left(1 + \dfrac{1}{l}\hat{\rho}_p^{(1)}\right)^2} \frac{d\sigma}{e-d} \frac{d\alpha_p^0}{\mathscr{M}_p}$$

$$= \frac{\dfrac{1}{3}\dfrac{e^3 - d^3}{e-d} + \left(\dfrac{\hat{\rho}_p^{(1)}}{l}\right)^2 \left(\dfrac{\mathscr{M}_p^2}{3} + (c_p - \hat{c}_p)^2\right)}{\left(1 + \dfrac{1}{l}\hat{\rho}_p^{(1)}\right)^2},$$

$$\mathscr{D}_1^p(\alpha_{\text{lse}}^p) = \int_{c_p - \mathscr{M}_p}^{c_p + \mathscr{M}_p} \frac{d\alpha}{2\mathscr{M}_p} \int_d^e \frac{\sigma^2}{l} \frac{d\sigma}{e-d} = \frac{e^2 + ed + d^2}{3l},$$

hence

$$\frac{\mathscr{D}_1^p(\hat{\alpha})}{\mathscr{D}_1^p(\alpha_{\text{lse}})} = \frac{1 + \dfrac{1}{\rho_p^{(1)}}\left(\dfrac{\hat{\rho}_p^{(1)}}{l}\right)^2 v_1}{\left(1 + \dfrac{1}{l}\hat{\rho}_p^{(1)}\right)^2}, \qquad v_1 = 1 + 3\left(\frac{\hat{c}_p - c_p}{\mathscr{M}_p}\right)^2.$$

We now compute

$$\mathscr{D}_2^p(\hat{\alpha}) = \sup_{\alpha,\sigma} \frac{\dfrac{\sigma^2}{l} + \left(\dfrac{\hat{\rho}_p^{(2)}}{l}\right)^2 (c_p - \alpha_p^0)^2}{\left(1 + \dfrac{\hat{\rho}_p^{(2)}}{l}\right)^2} = \frac{\dfrac{e^2}{l} + \left(\dfrac{\hat{\rho}_p^{(2)}}{l}\right)^2 (|\hat{c}_p - c_p| + \mathscr{M}_p)^2}{\left(1 + \hat{\rho}_p^{(2)}\dfrac{1}{l}\right)^2}.$$

On the other hand,

$$\mathscr{D}_2^p(\alpha_{\text{lse}}) = \sup_{\sigma} \frac{\sigma^2}{l} = \frac{e^2}{l},$$

hence

$$\frac{\mathscr{D}_2^p(\hat{\alpha}_p)}{\mathscr{D}_2^p(\alpha_{\text{lse}})} = \frac{1 + \dfrac{1}{\rho_p^{(2)}}\left(\dfrac{\hat{\rho}_p^{(2)}}{l}\right)^2 v_2}{\left(1 + \dfrac{1}{l}\hat{\rho}_p^{(2)}\right)^2}, \qquad v_2 = \left(1 + \frac{|c_p - \hat{c}_p|}{\mathscr{M}_p}\right)^2.$$

The theorem is proved. □

We have thus studied the theory of estimating regression parameters. This theory is based on the fact that in a certain narrow class of estimators the least-squares method is optimal (for normal regression this class is the class of unbiased estimators, and for general regression theory it is the class of linear unbiased estimators). It then turned out that in a class of biased estimators, better estimators than those arising from the least-squares method

may be constructed. Such nonlinear biased methods of estimation were obtained for estimating parameters of normal regression, while linear biased methods arise in the general model of regression estimation.

Estimation methods presented in this chapter can be utilized for regression estimation provided the density $P(x)$ is known and the regression is indeed a linear function in the parameters.

Chapter 6

A Method of Minimizing Empirical Risk for the Problem of Pattern Recognition

§1 A Method of Minimizing Empirical Risk

In the preceding three chapters the estimation of dependences was associated with the methods of estimating probability densities. The determination of the function which minimizes the expected risk

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \tag{6.1}$$

on the basis of the empirical data

$$x_1, y_1; \ldots; x_l, y_l \tag{6.2}$$

was reduced to estimating the density $\hat{P}(x, y)$ on the basis of the sample (6.2) and minimization of the functional

$$I_{\text{emp}}(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(x, y)\, dx\, dy.$$

As was mentioned in Chapter 2, this method of minimizing the risk (6.1) generally is not reasonable, because the problem of density estimation is a more difficult problem than the minimization of the expected risk. Only when a substantial prior information is available about the desired density $P(x, y)$, so that the function $P(x, y)$ can be defined up to its parameters, is this approach plausible. Methods of parametric statistics developed for this particular case were utilized in the preceding chapters.

However, in specific problems the structure of the density $P(x, y)$ is unknown. Thus the successful application of methods of parametric statistics hinges on the assumption that the hypothesized density structure corresponds to the true one.

Starting with this chapter, we shall study methods of estimating dependences which do not require density estimation. The basis for these methods is the principle of minimizing the empirical risk, according to which as the minimum point of the functional (6.1) one takes the minimum point of the empirical functional

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2, \tag{6.3}$$

constructed from a random independent sample (6.2). Let the minimum of functional (6.3) be attained for $F(x, \alpha_{\text{emp}})$. The problem is to establish when the obtained function $F(x, \alpha_{\text{emp}})$ is close to the function $F(x, \alpha_0)$ which minimizes (6.1) in $F(x, \alpha)$.

Above (Chapter 2, Section 6) we have associated this problem with the problem of the uniform convergence of the means to their mathematical expectations, i.e., with the situation when for any given value of deviation $\varkappa$ the inequality

$$P\left\{\sup_{\alpha} |I(\alpha) - I_{\text{emp}}(a)| > \varkappa\right\} < \eta \tag{6.4}$$

can be asserted.

Let (6.4) be satisfied. Then the inequality

$$P\{I(\alpha_{\text{emp}}) - I(\alpha_0) > 2\varkappa\} < \eta \tag{6.5}$$

is valid. In other words, if (6.4) holds, then with probability $1 - \eta$ the deviation of the function (solution) $F(x, \alpha_0)$ which is the best in the class $F(x, \alpha)$ from the function which yields a minimum for the empirical risk $F(x, \alpha_{\text{emp}})$ does not exceed $2\varkappa$.

Indeed, the condition (6.4) implies that with probability $1 - \eta$ the two inequalities

$$\begin{aligned} I(\alpha_{\text{emp}}) - I_{\text{emp}}(\alpha_{\text{emp}}) &< \varkappa, \\ I_{\text{emp}}(\alpha_0) - I(\alpha_0) &< \varkappa \end{aligned} \tag{6.6}$$

are simultaneously satisfied. Moreover, since α_{emp} and α_0 are the minimum points of $I_{\text{emp}}(\alpha)$ and $I(\alpha)$, the inequality

$$I_{\text{emp}}(\alpha_{\text{emp}}) \leq I_{\text{emp}}(\alpha_0) \tag{6.7}$$

is valid. The inequalities (6.6) and (6.7) yield that

$$I(\alpha_{\text{emp}}) - I(\alpha_0) < 2\varkappa. \tag{6.8}$$

And since the inequalities (6.6) are both fulfilled simultaneously with probability $1 - \eta$, so is (6.8). Consequently

$$P\{I(\alpha_{\text{emp}}) - I(\alpha_0) > 2\varkappa\} < \eta. \tag{6.9}$$

In this chapter we shall consider the theory of uniform convergence of the means to the mathematical expectations as applied to the problem of pattern recognition: i.e., in the case when the loss function in the functional of expected risk takes only two values, zero and one. In Chapter 7, for the problem of regression estimation we shall extend the results obtained to the case when the loss function takes on an arbitrary form in the interval $(0, \infty)$. It is important to note here that the validity of basic theorems proved in these chapters does not depend on the form of the loss function. Therefore in spite of a quadratic loss function used in the text we shall obtain a general theory of risk minimization.

§2 Uniform Convergence of Frequencies of Events to Their Probabilities

Consider the functional whose minimization is the essence of the pattern recognition problem:

$$I(\alpha) = P(\alpha) = \int (\omega - F(x, \alpha))^2 P(x, \omega)\, dx\, d\omega. \tag{6.10}$$

As has already been mentioned, this functional defines for each decision rule the probability of erroneous classification. The empirical functional

$$I_{\text{emp}}(\alpha) = \nu(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (\omega_i - F(x_i, \alpha))^2, \tag{6.11}$$

computed by means of the sample

$$x_i, \omega_1; \ldots; x_l, \omega_l, \tag{6.12}$$

defines for each decision rule the frequency of incorrect classification.

According to the classical theorems of probability theory the frequency of occurrence of an event converges to the probability of this event as the number of trials increases indefinitely. Formally this means that for any fixed α and $\varkappa$ the relation

$$\lim_{l \to \infty} P\{|P(\alpha) - \nu(\alpha)| > \varkappa\} = 0 \tag{6.13}$$

holds. However (cf. Chapter 2, Section 6), the condition (6.13) does not imply that the rule which minimizes (6.11) will yield a value of the functional (6.10) close to the minimal. For l sufficiently large the proximity between the solution obtained and the best one does follow from a stronger condition which stipulates that the equality

$$\lim_{l \to \infty} P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\} = 0 \tag{6.14}$$

is valid for any $\varkappa$. In this case we say that the *uniform convergence of frequencies of events to their probabilities over a class of events* $S(\alpha)$ is valid. Each event $S(\alpha^*)$ in the class $S(\alpha)$ is given by the decision rule $F(x, \alpha^*)$ as a set of pairs x, ω for which the equality $(\omega - F(x, \alpha^*))^2 = 1$ is satisfied.

Below we shall present conditions which assure uniform convergence of frequencies of events to their probabilities and at the same time determine the domain of applicability of the method of minimizing empirical risk. However, we first note that application of the method of minimizing the empirical risk does not guarantee a successful solution of the problem of estimating dependences. Here is an example of an algorithm for pattern recognition which minimizes the empirical risk but at the same time one cannot guarantee that the constructed decision rule will be close to the best in a given class: Elements of the sample are stored in memory, and each situation to be recognized is compared with the examples available in memory. If the situation at hand coincides with one of the examples it will be attributed to the class to which the example belongs. If, however no analogous example is available in memory, the situation is attributed to the first class. It is obvious that such a device cannot improve itself, since usually only a negligible fraction of the possible situations will correspond to the sample. At the same time, such a device classifies the elements of the sample without error, i.e., the algorithm minimizes the empirical risk down to zero.

Below we shall verify that this algorithm uses a set of decision rules which form a system of events over which uniform convergence does not hold.

§3 A Particular Case

When does the uniform convergence of frequencies to probabilities take place? Consider the simple case where the class of decision rules $F(x, \alpha)$ is finite, consisting of N rules:

$$F(x, \alpha_1), \ldots, F(x, \alpha_N).$$

An event A_i corresponds to each decision rule $F(x, \alpha_i)$ consisting of pairs x, ω such that $(\omega - F(x, \alpha_i))^2 = 1$. This defines a finite number N of events $A_1, \ldots, A_N$.

For each fixed event the law of large numbers is valid (the frequency converges to the probability as the number of trials increases indefinitely). One of the specific forms of this law is the Hoeffding inequality:

$$P\{|P(\alpha_i) - \nu(\alpha_i)| > \varkappa\} < 2\exp\{-2\varkappa^2 l\}. \tag{6.15}$$

We are however interested in uniform convergence, i.e., in the probability of simultaneous fulfillment of inequalities

$$|P(\alpha_i) - \nu(\alpha_i)| \leq \varkappa, \qquad i = 1, 2, \ldots, N.$$

This probability can be easily bounded from above if the probability of occurrence of each one of the inequalities (6.15) is assessed separately:

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} \leq \sum_{i=1}^{N} P\{|P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}.$$

Taking into account the inequality (6.15), we obtain

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} < 2N \exp\{-2\varkappa^2 l\}. \tag{6.16}$$

This inequality implies that for a finite number of events the uniform convergence of frequencies of occurrences of events to the corresponding probabilities is always valid, i.e., the limit

$$\lim_{l\to\infty} P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} = 0.$$

We now require that the probability of the realization of the event

$$\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\}$$

not exceed η, i.e., that the inequality

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} < \eta \tag{6.17}$$

will be fulfilled. It follows from the bound (6.16) that the inequality (6.17) is definitely satisfied if the quantities N, l, $\varkappa$, and η are connected by

$$2N \exp\{-2\varkappa^2 l\} = \eta. \tag{6.18}$$

If one solves Equation (6.18) for $\varkappa$, then for given N, l, and η an estimator of the maximal deviation of the frequencies from the corresponding probability in the class of events under consideration is obtained:

$$\varkappa = \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}}. \tag{6.19}$$

If, however we solve Equation (6.18) for l, then we obtain the size of the sample such that with probability at least $1 - \eta$ one can assert that the maximal deviation of the frequency from the probability over this class does net exceed $\varkappa$:

$$l = \frac{\ln N - \ln(\eta/2)}{2\varkappa^2}. \tag{6.20}$$

We have thus proved the following theorem:

Theorem 6.1. *Let the set of decision rules consist of N elements, and for decision rules $F(x, \alpha_i)$ let the frequency of errors in the sample of size l be equal to*

$\nu(\alpha_i)$. *Then with probability* $1 - \eta$ *one may assert that the inequality*

$$\nu(\alpha_i) - \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}} \leq P(\alpha_i) \leq \nu(\alpha_i) + \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}}$$

is valid simultaneously for all decision rules.

Remark. Since the inequalities are valid for all N rules, Theorem 6.1 determines a confidence interval for the quality of a decision rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk among N rules. This interval is

$$\nu(\alpha_{\text{emp}}) - \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}} \leq P(\alpha_{\text{emp}}) \leq \nu(\alpha_{\text{emp}}) + \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}}.$$

In what follows the upper bound will be of importance: with probability $1 - \eta$,

$$P(\alpha_i) \leq \nu(\alpha_i) + \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}}$$

is valid simultaneously for all decision rules (including those which minimize empirical risk).

§4 A Deterministic Statement of the Problem

The size of the confidence interval computed based on Theorem 6.1 may be excessive. Indeed, consider the case when the set consisting of N decision rules contains a rule which solves perfectly the problem of pattern recognition, i.e., a rule for which the possibility of erroneous classification will equal zero. Such a formulation of the problem is sometimes called *deterministic.*† Then this rule (or a rule close to it) should be found from the sample $x_1, \omega_1; \ldots; x_l, \omega_l$.

We seek this rule using the method of minimizing the empirical risk. Since there exists among functions $F(x, \alpha_i)$ $(i = 1, \ldots, N)$ a function which solves the problem perfectly, it is clear *a priori* that for any sample $x_1, \omega_1; \ldots; x_l, \omega_l$ the value of the minimum of empirical risk will be zero. This minimum, however, can be obtained for several functions. Thus it becomes necessary to estimate the probability that the quality of any function which yields a value of zero for the empirical risk will not be worse than the given $\varkappa$.

Introduce the function

$$\theta(z) = \begin{cases} 1 & \text{for } z = 0, \\ 0 & \text{for } z > 0. \end{cases}$$

† The terminology is unfortunate, since the problem remains statistical. However, we use it because it is widespread.

Then an estimate of the rate of uniform convergence of frequencies to probabilities over the set of events for which the frequency of errors is zero is an estimate of the probability of an event

$$\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)|\,\bar{\theta}(\nu(\alpha_i)) > \varkappa\right\}$$

(rather than the event $\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}$ as in Theorem 6.1).

Since the number of functions for which the zero value of empirical risk is attained does not exceed N (the total number of the functions in this class), the inequality

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)|\,\bar{\theta}(\nu(\alpha_i)) > \varkappa\right\} \leq NP_\varkappa \tag{6.21}$$

is valid. Here $P_\varkappa$ is the probability that the decision rule for which the probability of committing an error exceeding $\varkappa$ will classify correctly all the elements of the sample. This probability may be easily bounded:

$$P_\varkappa \leq (1 - \varkappa)^l. \tag{6.22}$$

Substituting the bound for $P_\varkappa$ into (6.21), we obtain

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)|\,\bar{\theta}(\nu(\alpha_i)) > \varkappa\right\} \leq N(1 - \varkappa)^l. \tag{6.23}$$

In order that the probability

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)|\,\bar{\theta}(\nu(\alpha_i)) > \varkappa\right\}$$

may not exceed the value η, it is sufficient that the equality

$$N(1 - \varkappa)^l = \eta \tag{6.24}$$

be fulfilled. Solving this equation with respect to l, we obtain

$$l = \frac{\ln N - \ln \eta}{-\ln(1 - \varkappa)}. \tag{6.25}$$

Since for small $\varkappa$ the approximation

$$-\ln(1 - \varkappa) \approx \varkappa$$

is valid, (6.25) may be represented in the form

$$l = \frac{\ln N - \ln \eta}{\varkappa}.$$

In contrast with (6.20), the denominator here is $\varkappa$ rather than $2\varkappa^2$, i.e., in the deterministic formulation the sufficient size of the sample is smaller than

in the general case. Solving (6.24) with respect to $\varkappa$, we obtain

$$\varkappa = \frac{\ln N - \ln \eta}{l}.$$

Thus the following theorem is valid:

Theorem 6.2. *If one chooses from the set of decision rules consisting of N elements a rule that commits no errors in the sample, then with probability $1 - \eta$ one can assert that the probability of erroneous classification using the selected rule is within the limits*

$$0 \leq P \leq \varkappa,$$

where

$$\varkappa = \frac{\ln N - \ln \eta}{l}.$$

§5 Upper Bounds on Error Probabilities

Despite their apparent simplicity, Theorems 6.1 and 6.2 are quite deep. Essentially the subsequent development of the theory of minimizing empirical risk consists of a generalization of these theorems to the case of infinitely many decision rules. The basic points of this further theory are already available. We shall dwell on them in some detail.

(1) Theorems 6.1 and 6.2 are immediately obtained from the bounds on the rate of uniform convergence, over a class of events, of frequencies to probabilities. Theorem 6.1 is based on the bound (6.16) on the rate of uniform convergence over the class of events S_N: $A_1, \ldots, A_N$ of frequencies towards probabilities. Theorem 6.2 is based on a bound on the rate of uniform convergence over a narrower class $\{|P(\alpha_i) - \nu(\alpha_i)| \bar{\theta}(\nu(\alpha_i)) \leq \varkappa\}$. Denote this class by $\hat{S}_N$.

(2) In both cases the rate of uniform convergence was determined by the product of two quantities: the number of events in a class, and a bound on the probability that the frequency of any fixed event in the class deviates by more than $\varkappa$ from the probability of this event. For the events considered in Theorem 6.1 this probability does not exceed $\exp\{-2\varkappa^2 l\}$; for the events considered in Theorem 6.2 the analogous probability does not exceed $(1 - \varkappa)^l \approx \exp\{-\varkappa l\}$. Thus a bound on the rate of uniform convergence of frequencies to probabilities over a class of events is obtained from a bound on the rate of the ordinary convergence which follows from the law of large numbers, by multiplying it by the number of events in this class. When constructing a theory of uniform convergence over a class of events with an infinite number of members, this structure of a bound on the rate of uniform

convergence is retained. However, instead of the number of events, in this case other characteristics of the "capacity" of the class of events are utilized.

(3) In Theorem 6.1 two-sided bounds on the probability of erroneous classification using a decision rule which minimizes the empirical risk were obtained. However, for the subsequent theory the lower bound is of little importance. Therefore it is of interest to obtain a bound on a uniform one-sided deviation, i.e., a bound on

$$P\left\{\sup_i (P(\alpha_i) - \nu(\alpha_i)) > \varkappa\right\},$$

and not on

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\}.$$

The probability of the event $\{\sup_i (P(\alpha_i) - \nu(\alpha_i)) > \varkappa\}$ does not exceed the probability of the event $\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}$. Consequently a more refined bound on the probability of a uniform one-sided deviation $P\{\sup_i (P(\alpha_i) - \nu(\alpha_i)) > \varkappa\}$, than that on the probability of a two-sided uniform deviation $P\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}$ is possible. Such a bound allows us to obtain from the above a bound on the probability of erroneous classification which is better than the one obtained from Theorem 6.1.

(4) The bounds on the rate of uniform convergence given by (6.16) and (6.23) depend substantially on bounds on the probability of deviation of a frequency from the probability of events in the class under consideration (S_N or $\hat{S}_N$). The least favorable event A for the class S_N is that for which $P(A) = \frac{1}{2}$. Therefore only the bound (6.16) is possible. For the class of events $\hat{S}_N$ the least favorable event is the one for which $P(A) = \varkappa$. The more refined bound (6.22) is available for the probability of deviation of the frequency from the probability of this event. Thus the bounds obtained for the classes of events S_N and $\hat{S}_N$ differ in the same manner as the bound on the probability of a deviation of an event A such that $P(A) = \frac{1}{2}$ differs from the corresponding bound on an event A' such that $P(A') = \varkappa$. This fact demands that more careful attention be given to the requirements imposed on the amounts of deviation of frequencies from the respective probabilities for different events in the class. For our purposes of obtaining a uniform bound on the risk it is reasonable not to require a uniform deviation of frequencies from probabilities for all events in the class but to allow a larger deviation for events such that $P(A)$ is close to $\frac{1}{2}$ and a smaller one for events such that $P(A')$ is close to $\varkappa$. For example, it makes sense to bound the uniform relative value of the deviation

$$\left\{\sup_i \frac{P(\alpha_i) - \nu(\alpha_i)}{\sigma(\alpha_i)} > \varkappa\right\},$$

where $\sigma(\alpha_i) = \sqrt{P(\alpha_i)(1 - P(\alpha_i))}$; for small $P(\alpha_i)$ the approximation $\sigma(\alpha_i) \approx \sqrt{P(\alpha_i)}$ is valid. We now obtain a bound on the probability of the

one-sided relative deviation

$$P\left\{\sup_i \frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} > \varkappa\right\}, \tag{6.26}$$

and using it we shall construct an upper bound on the probability of erroneous classification. To derive the bound (6.26) we shall utilize the inequality

$$P\left\{\frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} > \varkappa\right\} < \exp\{-\tfrac{1}{2}\varkappa^2 l\}. \tag{6.27}$$

It follows from (6.27) that for a class consisting of N events the following bound on the rate of uniform convergence is valid:

$$P\left\{\sup_i \frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} > \varkappa\right\} < N \exp\{-\tfrac{1}{2}\varkappa^2 l\}. \tag{6.28}$$

We shall require that the probability of uniform one-sided relative deviation (6.28) not exceed η:

$$N \exp\{-\tfrac{1}{2}\varkappa^2 l\} = \eta.$$

This is certainly satisfied if

$$\varkappa = \sqrt{2\frac{\ln N - \ln \eta}{l}}. \tag{6.29}$$

Let the condition (6.29) be fulfilled. Then the inequality

$$\frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} < \varkappa \tag{6.30}$$

is satisfied simultaneously for all events A_i with probability $1 - \eta$. Solving (6.30) for $P(\alpha_i)$, we obtain that

$$P(\alpha_i) < \frac{\varkappa^2}{2}\left(1 + \sqrt{1 + \frac{4\nu(\alpha_i)}{\varkappa^2}}\right) + \nu(\alpha_i) \tag{6.31}$$

is valid with probability $1 - \eta$ for all the events in the class simultaneously.

Substituting (6.29) into (6.31), we obtain that with probability $1 - \eta$, the N simultaneous inequalities

$$P(\alpha_i) \leq \frac{\ln N - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2\nu(\alpha_i) l}{\ln N - \ln \eta}}\right) + \nu(\alpha_i)$$

are fulfilled. We have thus proved the following theorem:

Theorem 6.3. *Let the set of decision rules consist of N elements, and for each rule $F(x, \alpha_i)$ let the frequency of errors in the sample equal $\nu(\alpha_i)$. Then one can*

assert with probability $1 - \eta$ *that the bounds*

$$P(\alpha_i) \leq \frac{\ln N - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2\nu(\alpha_i)l}{\ln N - \ln \eta}}\right) + \nu(\alpha_i) \qquad (6.32)$$

are fulfilled simultaneously for all decision rules in the class.

Remark. Since the bound (6.32) is valid, with probability $1 - \eta$, simultaneously for all the rules in the class, it also holds for the rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk.

Theorem 6.3 allows us to estimate the quality of the rule which minimizes the empirical risk. Moreover, the bound (6.32) coincides with the bound given in Theorem 6.2 obtained in the extreme case when $P(\alpha^*) \approx 0$, and it is close to the bound given in Theorem 6.1 for the second extreme case when $P(\alpha^*) \approx \frac{1}{2}$. The structure of bounds for an infinite class of decision rules is the same.

§6 An ε-net of a Set

In the preceding sections we established the existence of a uniform convergence of frequencies of occurrences of events to the corresponding probabilities over a class of events consisting of a finite number of elements; we obtained bounds on the rate of this convergence and using it, bounds on the quality of a decision rule which minimizes the empirical risk. Our task is to generalize these results to the case of infinitely many events.

In general, however, in the infinite case the uniform convergence of frequencies to probabilities may not occur: for example, if the set of events is defined as consisting of all open subsets of the set X, ω. In this case a situation may arise where (cf. the example in Section 2) an algorithm for minimizing the empirical risk yields the value zero for the risk but it is not capable of learning. Therefore the problem is to determine conditions which will assure uniform convergence for an infinite number of events, to bound its rate, and finally to obtain an upper bound on the probability of erroneous classification for a rule which minimizes the empirical risk.

In mathematics the necessity often arises of extending results valid for a finite set of elements to the infinite case. Usually such a generalization is possible if the infinite set can be covered by a *finite ε-net.*

Definition. The set B of elements in a metric space R is called an *ε-net* of the set G if any point $c \in G$ is distant from some point $b \in B$ by an amount not exceeding ε, i.e., $\rho(b, c) < \varepsilon$.

We say that the set G *admits a covering by a finite ε-net* if for each ε there exists an ε-net B consisting of a finite number of elements.

In this section, for an infinite set of decision rules admitting a covering by a finite ε-net we shall obtain assertions analogous to the assertions of Theorems 6.1 and 6.3.

Thus let an infinite set of decision rules $F(x, \alpha)$ be given on which the metric $\rho(\alpha_1, \alpha_2) = \rho(F(x, \alpha_1), F(x, \alpha_2))$ is defined and a finite ε-net is singled out. Let this finite ε-net consist of $N(\varepsilon)$ elements. Moreover, let it be given that if two decision rules $F(x, \alpha_1)$ and $F(x, \alpha_2)$ are distant from each other by an amount not exceeding ε ($\rho(\alpha_1, \alpha_2) \leq \varepsilon$), then the quality of these rules differs by an amount not exceeding $\delta(\varepsilon)$, i.e.,

$$\left| \int (\omega - F(x, \alpha_1))^2 P(x, \omega)\, dx\, d\omega - \int (\omega - F(x, \alpha_2))^2 P(x, \omega)\, dx\, d\omega \right| \leq \delta(\varepsilon).$$

In other words, a small variation in the decision rule implies a small variation in the quality of classification.

Under these conditions Theorems 6.1 and 6.3 can be generalized as follows:

Theorem 6.4. *Let the set of decision rules $F(x, \alpha)$ be covered by a finite ε-net. Then with probability $1 - \eta$ the quality of the decision rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk is bounded by*

$$\nu(\alpha_i(\alpha_{\text{emp}})) - \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} - \delta(\varepsilon) \leq P(\alpha_{\text{emp}})$$

$$\leq \nu(\alpha_i(\alpha_{\text{emp}})) + \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} + \delta(\varepsilon),$$

where $F(x, \alpha_i(\alpha_{\text{emp}}))$ is an element of the ε-net which is closest to $F(x, \alpha_{\text{emp}})$.

Theorem 6.5. *Let the set of decision rules $F(x, \alpha)$ be covered by a finite ε-net. Then with probability $1 - \eta$ the quality of the decision rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk is bounded by*

$$P(\alpha_{\text{emp}}) \leq \nu(\alpha_i(\alpha_{\text{emp}})) + \frac{\ln N(\varepsilon) - \ln \eta}{l} \left(1 + \sqrt{1 + \frac{2\nu(\alpha_i(\alpha_{\text{emp}}))l}{\ln N(\varepsilon) - \ln \eta}}\right) + \delta(\varepsilon),$$

where $F(x, \alpha_i(\alpha_{\text{emp}}))$ is an element of the ε-net which is closest to $F(x, \alpha_{\text{emp}})$.

Remark. Theorems 6.4 and 6.5 are valid for any ε-net given *a priori* (before the appearance of the sample). In particular the value of ε which defines the ε-net can be chosen in Theorem 6.4 from the condition of the minimum of expression

$$\sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} + \delta(\varepsilon),$$

and in Theorem 6.5 from the condition of the minimum of expression

$$\frac{\ln N(\varepsilon) - \ln \eta}{l} \left(1 + \sqrt{1 + \frac{2cl}{\ln N(\varepsilon) + \ln \eta}}\right) + \delta(\varepsilon),$$

where $0 \leq c \leq 1$ is a constant (for example $c = 0.5$).

Theorems 6.4 and 6.5 are proved in the same way:

PROOF.

(1) A finite ε-net consisting of $N(\varepsilon)$ elements

$$F(x, \alpha_1), \ldots, F(x, \alpha_{N(\varepsilon)}) \tag{6.33}$$

is given for the set of decision rules $F(x, \alpha)$. According to Theorem 6.1 (6.3) the inequalities

$$\begin{gathered} \nu(\alpha_i) - \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} \le P(\alpha_i) \le \nu(\alpha_i) + \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}}, \\ \left(P(\alpha_i) \le \frac{\ln N(\varepsilon) - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2\nu(\alpha_i)l}{\ln N(\varepsilon) - \ln \eta}}\right) + \nu(\alpha_i)\right) \end{gathered} \tag{6.34}$$

are fulfilled with probability $1 - \eta$ simultaneously for all $N(\varepsilon)$ elements of (6.33).

(2) For any decision rule $F(x, \alpha^*)$ (including the one which minimizes in $F(x, \alpha)$ the value of the empirical risk), the closest element of the ε-net $F(x, \alpha_i(\alpha^*))$ can be found, for which this element satisfies

$$|P(\alpha^*) - P(\alpha_i(\alpha^*))| \le \delta(\varepsilon). \tag{6.35}$$

The inequalities (6.34) and (6.35) imply that for the decision rule $F(x, \alpha_i(\alpha_{\text{emp}}))$ the relations

$$\begin{gathered} \nu(\alpha_i(\alpha_{\text{emp}})) - \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}} - \delta(\varepsilon) \\ \le P(\alpha_{\text{emp}}) \le \nu(\alpha_i(\alpha_{\text{emp}})) + \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} + \delta(\varepsilon), \\ \left(P(\alpha_{\text{emp}}) \le \frac{\ln N(\varepsilon) - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2\nu(\alpha_i(\alpha_{\text{emp}}))l}{N(\varepsilon) - \ln \eta}}\right) + \delta(\varepsilon) + \nu(\alpha_i(\alpha_{\text{emp}}))\right) \end{gathered}$$

are valid with probability $1 - \eta$. The theorems are thus proved. □

Thus if the set of decision rules $F(x, \alpha)$ admits a cover by a finite ε-net and the distribution $P(x, \omega)$ is such that close values of the probability of erroneous classification correspond to close decision rules, then as the sample size increases the method of minimizing the empirical risk should in principle successfully yield the desired result.† Moreover for each fixed ε the probability of erroneous classification using the rule which minimizes the empirical risk is bounded in terms of the inequalities (6.34).

However, in order to utilize these bounds the value of $\delta(\varepsilon)$ is required. To compute this value the density $P(x)$ is used, which in the formulation of the problem of pattern recognition is assumed to be unknown. In the next chapter, when solving the problem of estimating regression, we shall obtain the value of $\delta(\varepsilon)$ and be able to utilize bounds on the quality of a function expressed in terms of the value of empirical risk $\delta(\varepsilon)$ and $N(\varepsilon)$. In this chapter, to obtain the rate of uniform convergence of frequencies to the respective probabilities over an infinite class of events, a new idea will be utilized. This will eventually lead us to the construction of necessary and sufficient conditions for uniform convergence, to the derivation of a bound on the rate of uniform convergence based on these conditions, and finally to a constructive bound on the quality of a decision rule obtained using the method of minimizing the empirical risk.

† Although this assertion does not follow formally from Theorem 6.4, its proof is completely analogous.

§7 Necessary and Sufficient Conditions for Uniform Convergence of Frequencies to Probabilities

Up until now we have utilized quite rough "capacity" characteristics of the set of decision rules (the number of elements in the set) to obtain bounds on the rate of uniform convergence. In this section we introduce a more refined characteristic of capacity—*the entropy of a system of events on samples of size l.* Using this characteristic one can establish exhaustive necessary and sufficient conditions for uniform convergence of frequencies of events to their respective probabilities, i.e., for the equality

$$\lim_{l\to\infty} P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\} = 0$$

to be valid for any $\varkappa$.

Thus let a set S of decision rules $F(x, \alpha)$ be defined and a sample $x_1, \ldots, x_l$ be given. This sample can generally be subdivided into two classes in 2^l ways. However, only those subdivisions of the sample which can be accomplished using the rules $F(x, \alpha)$ will be of interest. (Uusing the rule $F(x, \alpha^*)$, the set $x_1, \ldots, x_l$ is subdivided into two subsets: one on which $F(x, \alpha^*) = 1$, and the other on which $F(x, \alpha^*) = 0$.) The number of different subdividing methods depends on the class of decision rules $F(x, \alpha)$ as well as on the sample. We shall denote this number by

$$\Delta^S(x_1, \ldots, x_l).$$

Consider the system of events

$$S(\alpha) = \{x, \omega: (\omega - F(x, \alpha))^2 = 1\}$$

formed by the set of decision rules $F(x, \alpha)$. Let a random independent sample

$$x_1, \omega_1; \ldots; x_l, \omega_l \tag{6.36}$$

be given. The system of events $S(\alpha)$ induces $\Delta(S(\alpha); x_1, \omega_1; \ldots; x_l, \omega_l)$ different subsamples on the sample (6.36). Clearly the number of these subsamples equals $\Delta^S(x_1, \ldots, x_l)$. Since $x_1, \ldots, x_l$ is a random independent sample the number of subdivisions $\Delta^S(x_1, \ldots, x_l)$ is a random variable.

Definition. The quantity

$$H^S(l) = M \ln \Delta^S(x_1, \ldots, x_l)$$

is called the *entropy* of a system of events $S(\alpha)$ on a sample of size l.

It turns out that for the uniform convergence of frequencies $\nu(\alpha)$ to the respective probabilities $P(\alpha)$ over the set of events, it is necessary and sufficient that as the sample size increases, the portion of the entropy due to

a single element of the sample approach zero, i.e., that the sequence

$$\frac{H^S(1)}{1}, \frac{H^S(2)}{2}, \ldots, \frac{H^S(l)}{l}$$

approach zero as l increases. In other words the condition

$$\lim_{l \to \infty} \frac{H^S(l)}{l} = 0 \tag{6.37}$$

should be fulfilled. The proof of this assertion follows from Theorem A.1 of the Appendix to Chapter 7.

Like any exhaustive conditions, the necessary and sufficient conditions stated above for the uniform convergence of frequencies to their respective probabilities utilize some refined concepts. In our case such a concept is the entropy $H^S(l)$ of a system of events $S(\alpha)$ on samples of size l, which is constructed by means of the density $P(x)$. In the case of the problem of pattern recognition the density is unknown, as stated above. Therefore, in order to establish the feasibility of minimizing the expected risk via the determination of the minimum of empirical risk, the necessary and sufficient conditions (6.37) cannot be used.

For this reason it is important to obtain less refined sufficient conditions which firstly will not depend on the properties of the measure $P(x)$ and secondly will admit a bound on the rate of uniform convergence. Such conditions may be stated in terms of a capacity measure of the system of events $S(\alpha)$ which is obtained from the entropy $H^S(l)$ by abstracting it from measure properties.

Definition. The function

$$m^S(l) = \max_{x_1, \ldots, x_l} \Delta^S(x_1, \ldots, x_l),$$

where the maximum is taken over all possible samples of size l, is called the *growth function* of a system of events formed by the decision rules $F(x, \alpha)$.

The growth function is constructed in such a manner that it does not depend on the properties of measure $P(x)$ and the inequality

$$\ln m^S(l) \geq H^S(l) \tag{6.38}$$

is always satisfied. Now if the quantity

$$\frac{\ln m^S(l)}{l}$$

approaches zero as l increases, then in view of (6.38) the ratio $H^S(l)/l$ tends to zero *a fortiori*. Therefore the condition

$$\lim_{l \to \infty} \frac{\ln m^S(l)}{l} = 0$$

is a sufficient condition for the uniform convergence of frequencies to their probabilities. Below we shall show that the growth function can be easily obtained for the events defined by various classes of decision rules $F(x, \alpha)$ and hence the uniform convergence can be established. Moreover, as will be shown below, the rate of uniform convergence can also be estimated using the growth function $m^S(l)$.

§8 Properties of Growth Functions

A growth function has a simple interpretation: it counts the maximal number of ways for subdividing l points into two classes using the decision rules $F(x, \alpha)$. For growth functions the following remarkable theorem is valid.

Theorem 6.6. *A growth function is either identically equal to 2^l or for $l > h$ is majorized by the function*

$$m^S(l) < 1.5 \frac{l^h}{h!},$$

where $h + 1$ is the smallest sample size such that the condition $m^S(l) = 2^l$ is violated. In other words

$$m^S(l) \begin{cases} \text{either} \equiv 2^l, \\ \text{or} \quad < 1.5 \dfrac{l^h}{h!} \quad (l > h). \end{cases}$$

The proof of this theorem is presented in the appendix to this chapter.

In order to bound a growth function it is necessary to show that either (1) for any l, points $x_1, \ldots, x_l$ exist such that using the decision rules $F(x, \alpha)$ it would be possible to subdivide them into two classes by any one of the 2^l ways, or (2) a number h exists such that h points can be subdivided into classes in all possible ways, but $h + 1$ points cannot. In the first case the growth function is exponential; in the second it is polynomial. The number h can serve as the measure of diversity of the class of decision rules.

Definition. We say that the class of indicator functions *has capacity h* if the inequality

$$m^S(l) < 1.5 \frac{l^h}{h!} \qquad (l > h) \tag{6.39}$$

is valid. If the equality

$$m^S(l) \equiv 2^l$$

is satisfied we say that the capacity h of the class of indicator functions $F(x, \alpha)$ is *infinite*.

It is easy to verify that if the capacity of the class of indicator functions is finite, then the uniform convergence of frequencies to the respective probabilities always occurs. Indeed, in this case the relation

$$0 \leq \lim_{l \to \infty} \frac{\ln m^S(l)}{l} \leq \lim_{l \to \infty} \frac{h \ln l - \sum_{i=1}^{h} \ln i}{l} = 0$$

is valid and the sufficient condition is fulfilled.

The following class of decision rules, which are linear in the parameter, plays an important role in the subsequent theory:

$$F(x, \alpha) = \theta\left(\sum_{i=1}^{n} \alpha_i \varphi_i(x)\right); \qquad \theta(z) = \begin{cases} 1 & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases} \tag{6.40}$$

It is easy to obtain a growth function for a class of events defined by linear decision rules (6.40). For this purpose it is sufficient to determine the maximal number h of points in the space of dimensionality n which can be subdivided into two classes using a hyperplane in any one of the 2^h ways. It is known that this number equals n. Therefore according to Theorem 6.6 the growth function is bounded by

$$m^S(l) < 1.5 \frac{l^n}{n!} \qquad (l > n)$$

for the class of linear decision rules (6.40). Consequently for the class of linear decision rules sufficient conditions for uniform convergence are fulfilled.

It was shown in Chapter 2 that uniform convergence of frequencies of events to their probabilities over a class of events defined by one-dimensional linear decision rules $F(x, \alpha) = \theta(x + \alpha)$ makes up the content of the Glivenko–Cantelli theorem, which asserts the uniform convergence of the empirical cumulative distribution function to the population one.

§9 Bounds on Deviations of Empirically Optimal Decision Rules

In the appendix to this chapter a bound on the rate of uniform convergence of frequencies to probabilities over a class of events $S(\alpha)$ is obtained. It is shown that the inequality

$$P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\} < 6m^S(2l) \exp\left\{-\frac{\varkappa^2 l}{4}\right\} \tag{6.41}$$

is valid. The bound (6.41) is of the same form as the above: it is formed by multiplying the quantity $6m^S(2l)$—which is the capacity characteristic of the

system of events—by a bound on the probability that the deviation of the frequency from its probability exceeds $\varkappa$ (the quantity $\exp\{-\varkappa^2 l/4\}$).

If the capacity of the class of decision rules is infinite ($m^S(l) \equiv 2l$), then the bound (6.41) is trivial, since for all $\varkappa$ the right-hand side of the inequality exceeds 1. The bound (6.41) is meaningful when the capacity of the class of decision rules is finite:

$$m^S(l) < 1.5\frac{l^h}{h!}.$$

In this case it takes the form

$$P\left\{\sup_\alpha |P(\alpha) - \nu(\alpha)| > \varkappa\right\} < 9\frac{(2l)^h}{h!}\exp\left\{-\frac{\varkappa^2 l}{4}\right\}. \tag{6.42}$$

As l increases, the right-hand side of the inequality (6.42) tends to zero and the approach is faster for smaller values of the capacity h. We shall require that the probability

$$P\left\{\sup_\alpha |P(\alpha) - \nu(\alpha)| > \varkappa\right\}$$

not exceed η. This is certainly true if

$$9\frac{(2l)^h}{h!}\exp\left\{-\frac{\varkappa^2 l}{4}\right\} = \eta. \tag{6.43}$$

Equation (6.43) can be solved for κ (using Stirling's formula):

$$\varkappa = 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}}. \tag{6.44}$$

Then (6.42)–(6.44) imply the following theorem:

Theorem 6.7. *Let $F(x, \alpha)$ be the class of decision rules of bounded capacity h, and let $\nu(\alpha)$ be the frequency of errors computed from the sample for the rule $F(x, \alpha)$. Then with probability $1 - \eta$ one may assert that for $l > h$, and simultaneously for all decision rules $F(x, \alpha)$, the probability of erroneous classification is within the limits*

$$\nu(\alpha) - 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}} < P(\alpha) < \nu(\alpha) + 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}}.$$

Remark. Theorem 6.7 implies that for the rule $F(x, \alpha_{\text{emp}})$, which minimizes the empirical risk, the upper bound

$$P(\alpha_{\text{emp}}) < \nu(\alpha_{\text{emp}}) + 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}} \qquad (l > h)$$

is valid with probability $1 - \eta$.

In the appendix to this chapter it is shown that along with (6.41) the bound

$$P\left\{\sup_{\alpha} \frac{P(\alpha) - \nu(\alpha)}{\sqrt{P(\alpha)}} > \varkappa\right\} < 8m^s(2l)e^{-\varkappa^2 l/4}$$

is valid. This bound is nontrivial for a class of decision rules of bounded capacity:

$$P\left\{\sup_{\alpha} \frac{P(\alpha) - \nu(\alpha)}{\sqrt{P(\alpha)}} > \varkappa\right\} < 12\frac{(2l)^h}{h!}e^{-\varkappa^2 l/4}. \tag{6.45}$$

We shall require that the right-hand side of the inequality be equal to η:

$$12\frac{(2l)^h}{h!}e^{-\varkappa^2 l/4} = \eta.$$

This is fulfilled if

$$\varkappa = 2\sqrt{\frac{\ln\frac{(2l)^h}{h!} - \ln\frac{\eta}{12}}{l}} \approx 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{12}}{l}}. \tag{6.46}$$

On the other hand, the inequality (6.45) can be stated as follows: with probability η, simultaneously for all α the inequality

$$P(\alpha) \leq \frac{\varkappa^2}{2}\left(1 + \sqrt{1 + \frac{4\nu(\alpha)}{\varkappa^2}}\right) + \nu(\alpha) \tag{6.47}$$

is valid. The relations (6.46) and (6.47) imply the following theorem.

Theorem 6.8. *Let $F(x, \alpha)$ be a class of decision rules of bounded capacity h, and for each rule $F(x, \alpha)$ let the frequency of errors computed in the sample equal $\nu(\alpha)$. Then with probability $1 - \eta$ one can assert that the bound*

$$P(\alpha) \leq 2\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{12}}{l}\left(1 + \sqrt{1 + \frac{\nu(\alpha)l}{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{12}}}\right) + \nu(\alpha_{\text{emp}}) \tag{6.48}$$

is valid for $l > h$ simultaneously for all rules in the class.

Remark. It follows from Theorem 6.8 that for the rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk the bound

$$P(\alpha_{\text{emp}}) \leq 2\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{12}}{l}\left(1+\sqrt{1+\frac{\nu(\alpha_{\text{emp}})l}{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{12}}}\right)+\nu(\alpha_{\text{emp}})$$

is valid.

§10 Remarks on the Bound on the Rate of Uniform Convergence of Frequencies to Probabilities

In this chapter we have obtained bounds on the rate of uniform convergence of frequencies to the respective probabilities:

$$P\left\{\sup_{\alpha}|P(\alpha)-\nu(\alpha)|>\varkappa\right\}<\begin{cases}2Ne^{-2\varkappa^2 l},\\ 6m^s(2l)e^{-\varkappa^2 l/4},\end{cases}$$

and bounds on the uniform one-sided relative deviations of frequencies from their probabilities:

$$P\left\{\sup_{\alpha}\frac{P(\alpha)-\nu(\alpha)}{\sqrt{P(\alpha)}}>\varkappa\right\}<\begin{cases}Ne^{-\varkappa^2 l/2},\\ 8m^s(2l)e^{-\varkappa^2 l/4}\end{cases}$$

Using these bounds, Theorems 6.1, 6.3, 6.7, and 6.8 were obtained, which allow us to estimate the quality of a decision rule minimizing the empirical risk.

All the estimates obtained have the same structure, consisting of two factors: one which bounds the probability of the corresponding deviation (separately) for each event in the class, and another which characterizes the variety of the class of decision rules. Different characteristics of the variety of the class of decision rules are used for the bounds. The simplest is the number of decision rules in the class. The simplicity of this characteristic is due to the fact that it does not, for example, take into account whether the decision rules in the class are "substantially different" or whether all the rules are "equivalent."

An adequate measure of the variety of the class of decision rules, by which it is possible to construct necessary and sufficient conditions for the uniform convergence of frequencies to their probabilities, is the entropy of the system of events defined by the decision rules. However, to compute the entropy of a system of events on samples of length l is possible only if the density $P(x)$ is known, and it is assumed to be unknown in the formulation of the pattern recognition problem. Therefore a new measure of variety was introduced which is obtained from entropy by choosing the least favorable distri-

bution. This measure is expressed in terms of the capacity of the class of decision rules and can easily be computed.

Various definitions of measures of variety of a class of decision rules generate different theorems on the quality of algorithms minimizing the empirical risk. However, in all these theorems the very same fact is asserted: if the measure of variety of a class of decision rules is small compared with the sample size, then the method of minimizing empirical risk allows us to choose a rule which is close to the best one in the class.

A characteristic feature of the theory of minimizing empirical risk presented above is the complete absence of any indications as to the constructive feasibility of determining an algorithm. This feature has negative as well as positive aspects. On one hand, the theory does not give regular procedures for minimizing empirical riks; they should be implemented by a corresponding program. On the other hand, the theory is quite general. The method can be applied to various classes of decision rules: linear discriminant functions, piecewise linear discriminant functions, logistic functions of a particular kind, and so on. This is due to the fact that the theory of the method of minimizing empirical risk answers the question "what to do," leaving the question "how to do it" unsettled. Therefore various methods can be applied, including heuristic ones.

The application of heuristic methods in this case has some theoretical justification: if in a class of decision rules whose capacity is small compared to the sample size one chooses a rule which, while it does not yield the minimum of the empirical risk, results in a sufficiently small value of it, then in view of the theorems proved above, the decision rule selected will be of sufficiently high quality.

Constructive ideas for such algorithms admit a simple geometric interpretation: It is required to construct in a space X a hypersurface belonging to a given class of hypersurfaces which—with the smallest possible number of errors—will separate the vectors of the sample in one class from the corresponding vectors in the other. The assignment of vectors (including those which do not belong to a learning sequence) to a particular class is carried out according to the side of the subdividing hypersurface on which the vector is located.

Methods of constructing separating hypersurfaces constitute a constructive part of the theory of pattern recognition. These methods are presented in Addendum I.

§11 Remark on the General Theory of Uniform Estimating of Probabilities

We have thus developed a theory of uniform estimating of error probabilities in pattern recognition for arbitrary classes of decision rules. Formally, in the functional which computes the probabilities of errors we wrote a quadratic

loss function. In proving the related theorems, however, the form of the loss function was unimportant. What is important is that $Q(z, \alpha)$, $\alpha \in \Lambda$, is a class of indicator functions.

In fact, this chapter presents a theory more general than the uniform estimation of error probabilities in pattern recognition. Here a general theory has been developed for uniform estimation of probabilities from their frequencies in a class of events of limited capacity. We now formulate the basic assertions of this theory. The proofs are identical to those of similar theorems given in the chapter.

Assume that a space Z is given on which a probability measure $P(z)$ has been defined and a system of events S_α, $\alpha \in \Lambda$, is specified (subsets measurable with respect to the given measure and belonging to Z). Let $Q(z, \alpha)$, $\alpha \in \Lambda$ be a family of indicator functions on the sets S_α, $\alpha \in \Lambda$ (i.e., the function

$$Q(z, \alpha) = \begin{cases} 0 & \text{if } z \notin S_\alpha \\ 1 & \text{if } z \in S_\alpha \end{cases}.$$

Let the capacity of the family of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$, be finite and equal to h (there exists such an h that $m^{S_\alpha}(h) = 2^h$, $m^{S_\alpha}(h + 1) \neq 2^{h+1}$).

Under these conditions the following assertions hold on two-sided and one-sided uniform bounds of probabilities

$$P(\alpha) = \int_{S_\alpha} dP(z) = \int Q(z, \alpha)\, dP(z)$$

by virtue of associated frequencies

$$\nu(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha)$$

computed on a sample

$$z_1, \ldots, z_l.$$

Assertion 1. *For any* $l > (\Delta/(\Delta - 1))^2$, $\Delta > 1$ *with probability* $1 - \eta$ *simultaneously for all events* S_α, $\alpha \in \Lambda$, *the two-sided bound*

$$\nu(\alpha) - \Delta\sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{9}}{l}} \leq P(\alpha) \leq \nu(\alpha) + \Delta\sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{9}}{l}}$$

holds.

Assertion 2. *With probability* $1 - \eta$ *simultaneously for all events* S_α, $\alpha \in \Lambda$, *the one-sided bound*

$$P(\alpha) \leq \nu(\alpha) + 2\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{12}}{l}\left(1 + \sqrt{1 + \frac{\nu(\alpha)l}{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{12}}}\right)$$

holds.

Appendix to Chapter 6

Theory of Uniform Convergence of Frequencies to Probabilities: Sufficient Conditions†

§A1 Sufficient Conditions for Uniform Convergence of Frequencies to Probabilities

According to Bernoulli's classical theorem the frequency of occurrence of a certain event A in a sequence of independent trials converges (in probability) to the probability of this event. Often, however, it becomes necessary to assess simultaneously the probabilities of a class of events S based on the very same sample. Moreover, it is required that the frequencies converge to the probabilities uniformly over all events in the class S. More precisely, the probability that the maximal deviation over the class of frequencies from probabilities exceeds a given, arbitrarily small positive constant must tend to zero as the number of trials increases indefinitely.

It turns out that even in the simplest cases uniform convergence may not occur. Therefore a criterion is required which will test whether such convergence is present.

Let X be a set of elementary events on which a probability measure $P(x)$ is defined. Let S be a collection of random events, i.e., subsets of a space measurable with respect to the measure $P(x)$ (S is included in the σ-algebra of random events, but does not necessarily coincide with it). Denote by $X(l)$ the space of random independent samples taken from X of length l.

For each sample $X^l = x_1, \ldots, x_l$ and event $A \in S$, the frequency of occurrence of event A is defined as the ratio of the number $n(A)$ of elements of the sample belonging to A to the common sample size l:

$$\nu^l(A) = \nu(x_1, \ldots, x_l) = \frac{n(A)}{l}.$$

† Necessary and sufficient conditions for uniform convergence of frequencies to probabilities will follow from the results presented in the Appendix to Chapter 7.

Bernoulli's theorem asserts that for a fixed event A the deviation of the frequency from the probability tends to zero (in probability) with increasing sample size, i.e., for any $\varkappa$

$$P\{|P(A) - \nu^l(A)| > \varkappa\} \underset{l\to\infty}{\longrightarrow} 0.$$

Here, however, we are concerned with the maximal (over the class S) deviation of the frequency from the probability:

$$\pi(l) = \sup_{A\in S} |\nu^l(A) - P(A)|.$$

The quantity $\pi(l)$ is a function of a point in the space $X(l)$. We shall assume that this function is measurable with respect to a measure in $X(l)$, i.e., $\pi(l)$ is a random variable. The theorems below deal with bounds on the probabilities of the event $\pi(l)$.

§A2 The Growth Function

Let X be a set, S be a system of its subsets, and $X^l = x_1, \ldots, x_l$ be a sequence of elements x of length l. Each set $A \in S$ determines a subsequence X_A of this sequence consisting of elements belonging to A. We say that A induces a subsequence X_A on the sequence X^l.

Denote by

$$\Delta^S(x_1, \ldots, x_l)$$

the number of different subsequences X_A induced by the sets $A \in S$. Clearly,

$$\Delta^S(x_1, \ldots, x_l) \leq 2^l.$$

The number $\Delta^S(x_1, \ldots, x_l)$ is called the *index of the system S relative to the sample* $x_1, \ldots, x_l$.

The index of a system may be defined in another way as well. We shall consider $A_1 \in S$ to be equivalent to $A_2 \in S$ relative to the sample $x_1, \ldots, x_l$ if $X_{A_1} = X_{A_2}$. Then the index $\Delta^S(x_1, \ldots, x_l)$ is the number of equivalence classes into which the system S is subdivided by this equivalence relation.

Clearly the two definitions are equivalent. The function

$$m^S(l) = \max_{x_1,\ldots,x_l} \Delta^S(x_1, \ldots, x_l), \tag{A.1}$$

where the maximum is taken over all the sequences of length l is called *the growth function of the system S*. Here the maximum is always attained, since the index $\Delta^S(x_1, \ldots, x_l)$ takes on a finite number of values.

The growth function of a class of events possesses the following remarkable property.

Theorem A.1. *The growth function either is identically equal to 2^l or is bounded by the function*

$$\sum_{i=0}^{n-1} C_l^i$$

where n is the minimal value of l such that

$$m^S(l) \neq 2^l.$$

In other words

$$m^S(l) \begin{cases} \text{either} & \equiv 2^l, \\ \text{or} & < \sum_{i=0}^{n-1} C_l^i \end{cases} \tag{A.2}$$

To prove this assertion the following lemma is required.

Lemma A.1. *If for some sequence $x_1, \ldots, x_l$ and some n*

$$\Delta^S(x_1, \ldots, x_l) > \sum_{i=0}^{n-1} C_l^i,$$

then there exists a subsequence X^n of length n such that

$$\Delta^S(X^n) = 2^n.$$

PROOF. Denote

$$\sum_{i=0}^{n-1} C_l^i = \Phi(n, l)$$

(here and below we shall assume that $C_l^i = 0$ for $i > l$). For this function, as it is easy to verify, the relations

$$\begin{aligned} &\Phi(1, l) = 1, \\ &\Phi(n, l) = 2^l \quad \text{if } l \leq n + 1, \\ &\Phi(n, l) = \Phi(n, l-1) + \Phi(n-1, l-1), \quad \text{if } n \geq 2, l \geq 1 \end{aligned} \tag{A.3}$$

are valid. In turn these relations uniquely determine the function $\Phi(n, l)$ for $l > 0$ and $n > 0$.

We shall prove the lemma by an induction on l and n. For $n = 1$ and any $l \geq 1$ the assertion of the lemma is obvious. Indeed, in this case

$$\Delta^S(x_1, \ldots, x_l) > 1$$

implies that an element of the sequence x_i exists such that for some $A^* \in S$ we have $x_i \in A^*$, while for some other $A^{**} \in S$ we have $x_i \notin A^{**}$. Consequently,

$$\Delta^S(x_i) = 2.$$

For $l < n$ the assertion of the lemma is valid because the premise is false. Indeed, in this case the premise is

$$\Delta^S(x_1, \ldots, x_l) > 2^l,$$

which is impossible, since

$$\Delta^S(x_1, \ldots, x_l) \leq 2^l.$$

Finally assume that the lemma is valid for $n \leq n_0$ ($n_0 \geq 1$) for all l. Consider now the case $n = n_0 + 1$. We show that the lemma is valid in this case also for all l.

We fix $n = n_0 + 1$ and carry out the induction on l. As was pointed out, for $l < l_0 + 1$ the lemma is valid. We shall assume that it is valid for $l \leq l_0$ and show that it is valid for $l = l_0 + 1$. Indeed, let the condition of the lemma,

$$\Delta^S(x_1, \ldots, x_{l_0}, x_{l_0+1}) > \Phi(n_0 + 1, l_0 + 1)$$

be fulfilled for some sequence $x_1, \ldots, x_{l_0}, x_{l_0+1}$. The lemma will be proved if we find a subsequence of length $n_0 + 1$, say $X^{n_0+1} = x_1, \ldots, x_{n_0+1}$, such that

$$\Delta^S(x_1, \ldots, x_{n_0+1}) = 2^{n_0+1}.$$

Consider the subsequence $X^{l_0} = x_1, \ldots, x_{l_0}$. Two cases are possible:

(a) $\Delta^S(x_1, \ldots, x_{l_0}) > \Phi(n_0 + 1, l_0)$,
(b) $\Delta^S(x_1, \ldots, x_{l_0}) \leq \Phi(n_0 + 1, l_0)$.

In case (a), in view of the induction assumption, there exists a subsequence of length $n_0 + 1$ such that $\Delta^S(X^{n_0+1}) = 2^{n_0+1}$, q.e.d.

In case (b) we subdivide subsequences of the sequence X^{l_0} induced by the sets in S into two types. We assign to the first type subsequences X^r such that on the whole sequence X^{l_0+1} events belonging to S induce X^r as well as (X^r, x_{l_0+1}). Sequences X^r such that either X^r or (X^r, x_{l_0+1}) is induced on the sequence X^{l_0+1} are assigned to the second type. Denote the number of subsequences of the first type by K_1 and of the second by K_2. It is easy to see that

$$\Delta^S(x_1, \ldots, x_{l_0}) = K_1 + K_2,$$

$$\Delta^S(x_1, \ldots, x_{l_0}, x_{l_0+1}) = 2K_1 + K_2;$$

and hence

$$\Delta^S(x_1, \ldots, x_{l_0}, x_{l_0+1}) = \Delta^S(x_1, \ldots, x_{l_0}) + K_1. \tag{A.4}$$

Denote by S' the system of all subsets $A \in S$ that induce subsequences of the first type on the sequence X^{l_0}. Then if

(b') $K_1 = \Delta^{S'}(x_1, \ldots, x_{l_0}) > \Phi(n_0, l_0)$,

in view of the induction assumption there exists a subsequence $X^{n_0} = x_{i_1}, \ldots, x_{i_{n_0}}$ such that

$$\Delta^{S'}(x_{i_1}, \ldots, x_{i_{n_0}}) = 2^{n_0} \qquad (X^{n_0} \subset X^{l_0}).$$

However, in that case we have

$$\Delta^{S'}(x_1, \ldots, x_{i_{n_0}}, x_{l_0+1}) = 2^{n_0+1}$$

for the sequence $x_{i_1}, \ldots, x_{i_{n_0}}, x_{l_0+1}$, since for each subsequence X^r induced on the sequence X^{n_0}, two subsequences induced on X^r, x_{l_0+1} can be found, namely X^r and (X^r, x_{l_0+1}). Thus the required subsequence is obtained in case (b).

If, however

(b″) $K_1 = \Delta^{S'}(x_1, \ldots, x_{l_0}) \leq \Phi(n_0, l_0)$,

we then obtain in view of (A.4) and (b)

$$\Delta^{S}(x_1, \ldots, x_{l_0+1}) \leq \Phi(n_0 + 1, l_0) + \Phi(n_0, l_0),$$

which by virtue of the properties (A.3) of the function $\Phi(n, l)$ implies that

$$\Delta^{S}(x_1, \ldots, x_{l_0+1}) \leq \Phi(n_0 + 1, l_0 + 1).$$

This however contradicts the condition of the lemma (i.e., (b″) is impossible).

The lemma is proved. □

We shall now prove the theorem. As was pointed out, $m^S(l) \leq 2^l$. Let $m^S(l)$ not be identically equal to 2^l, and let n be the first value of l such that $m^S(l) \neq 2^l$. Then for any sample of size l larger than n, the inequality

$$\Delta^{S}(x_1, \ldots, x_l) \leq \Phi(n, l)$$

is valid. Indeed, otherwise, in view of the lemma's assertion, one could find a subsample $x_1, \ldots, x_n$ such that

$$\Delta^{S}(x_1, \ldots, x_n) = 2^n, \tag{A.5}$$

which is impossible, since by assumption $m^S(n) \neq 2^n$.

Thus the function $m^S(l)$ either is identically equal to 2^l or is majorized by $\Phi(n, l)$. The theorem is proved. □

Remark. The function $\Phi(n, l)$ can be bounded from the above for $n \leq 1$ and $l > n$ as follows:

$$\Phi(n, l) < 1.5 \frac{l^{n-1}}{(n-1)!}. \tag{A.6}$$

Since the relation (A.3) is fulfilled for $\Phi(n, l)$, to prove (A.6) it is sufficient to verify that for $n \geq 1$ and $l > n$ the inequality

$$\frac{l^{n-1}}{(n-1)!} + \frac{l^n}{n!} \leq \frac{(l+1)^n}{n!} \tag{A.7}$$

is valid and to verify (A.6) on the boundary, i.e., for $n = 1$ and $l = n + 1$.

The inequality (A.7) is clearly equivalent to

$$l^{n-1}(l + n) - (l + 1)^n \leq 0,$$

whose validity follows from Newton's binomial expansion.

It thus remains to verify A.6 on the boundary. For $n = 1$ the verification is direct. Next we shall verify the bound for small values of n and l:

$l = n + 1$	2	3	4	5	6
$\Phi(n, l)$	1	4	11	26	57
$1.5\dfrac{l^{n-1}}{(n-1)!}$	1.5	4.5	12	31.25	81

To check (A.6) for $n \geq 6$ we shall utilize Stirling's formula for an upper bound on $l!$:

$$l! \leq \sqrt{2\pi l}\, l^l e^{-l+(12l)^{-1}},$$

whence for $l = n + 1$

$$\frac{l^{n-1}}{(n-1)!} = \frac{(l-1)l^{(l-1)}}{l!} \geq \frac{l-1}{\sqrt{2\pi l}\, l}\, e^{-l+(12l)^{-1}},$$

and furthermore for $l \geq 6$

$$\frac{l^{(n-1)}}{(n-1)!} \geq 0.8\,\frac{1}{\sqrt{2\pi l}}\, e^l.$$

On the other hand, $\Phi(n, l) \leq 2^l$ always. Therefore it is sufficient to verify that for $l \geq 6$

$$2^l \leq 1.2\,\frac{1}{\sqrt{2\pi l}}\, e^l.$$

Actually it is sufficient to verify the inequality for $l = 6$ (which is carried out directly) since as l increases the right-hand side of the inequality grows faster than the left-hand side (for $l > 2$).

Thus we have seen that either the growth function is identically 2^l, or for some n the equality is violated for the first time (i.e., $m^S(n) \neq 2^n$), and then the growth function is bounded by a polynomial function

$$m^S(l) < 1.5\,\frac{l^{n-1}}{(n-1)!}.$$

Therefore in order to estimate the behavior of a growth function it is sufficient to find the smallest n such that on no sequence of length l does the system S induce all possible subsequences.

§A3 The Basic Lemma

Let a sample of size $2l$ be chosen:

$$X^{2l} = x_1, \ldots, x_l, x_{l+1}, \ldots, x_{2l},$$

and the frequencies of occurrence of the event $A \in S$ on the first half sample $x_1, \ldots, x_l$ and on the second half sample $x_{l+1}, \ldots, x_{2l}$ be computed. Denote these frequencies by $\nu'(A)$ and $\nu''(A)$ respectively, and consider the deviations of these quantities:

$$\rho_A(x_1, \ldots, x_{2l}) = |\nu'(A) - \nu''(A)|.$$

We are interested in the maximal deviation of the frequencies over all events of the class S:

$$\rho^S(x_1, \ldots, x_{2l}) = \sup_{A \in S} \rho_A(x_1, \ldots, x_{2l}).$$

Introduce the notation

$$\pi^S(x_1, \ldots, x_{2l}) = \sup_{A \in S} |\nu'(A) - P(A)|.$$

Furthermore we shall assume that $\pi^S(x_1, \ldots, x_l)$ and $\rho^S(x_1, \ldots, x_{2l})$ are measurable functions.

The Basic Lemma. *The distributions of the quantities* $\pi^S(x_1, \ldots, x_l)$ *and* $\rho^S(x_1, \ldots, x_{2l})$ *are related as follows*:

$$P\{\pi^S(x_1, \ldots, x_l) > \varkappa\} \leq 2P\left\{\rho^S(x_1, \ldots, x_{2l}) > \frac{\varkappa}{2}\right\},$$

provided that $l > 2/\varkappa$.

PROOF. By definition

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X(2l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X^{2l}),$$

where

$$\theta(z) = \begin{cases} 1 & \text{if } z > 0, \\ 0 & \text{if } z \leq 0. \end{cases}$$

Taking into account that the space $X(2l)$ of samples of size $2l$ is a direct product of $X_1(l)$ and $X_2(l)$ of half samples of size l, we have the equality

$$\int_{X(2l)} \varphi(x_1, \ldots, x_{2l})\, dX^{2l} = \int_{X_1(l)} \left[\int_{X_2(l)} \varphi(x_1, \ldots, x_{2l})\, dX_2^l\right] dX_1^l$$

for any measurable function $\varphi(x_1, \ldots, x_{2l})$, by Fubini's theorem [28].

Therefore

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X_1(l)} dP(X_1^l) \int_{X_2(l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l)$$

(in the inner integral the first half sample is fixed). Denote by Q the event in the space $X_1(l)$

$$\{\pi^S(x_1, \ldots, x_l) > \varkappa\},$$

and bounding the domain of integration, we obtain

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} \geq \int_Q dP(X_1^{2l}) \int_{X_2(l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l). \quad \text{(A.8)}$$

We now bound the inner integral on the right-hand side of the inequality and denote it by I. Here the sample $x_1, \ldots, x_l$ is fixed and is such that

$$\pi^S(x_1, \ldots, x_l) > \varkappa.$$

Consequently there exists an $A^* \in S$ such that

$$|P(A^*) - \nu(A^*; x_1, \ldots, x_l)| > \varkappa.$$

Then

$$I = \int_{X_2(l)} \theta\left[\sup_{A \subset S} \rho_A(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l) \geq \int_{X_2(l)} \theta\left[\rho_{A^*}(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l).$$

Let, for example,

$$\nu'(A^*; x_1, \ldots, x_l) < P(A^*) - \varkappa$$

(the case $\nu'(A^*) \geq P(A^*) + \varkappa$ is dealt with completely analogously). Then in order that the conditions

$$|\nu'(A^*; x_1, \ldots, x_l) - \nu''(A^*; x_{l+1}, \ldots, x_{2l})| > \frac{\varkappa}{2}$$

may be satisfied, it is sufficient that the relation

$$\nu''(A^*) > P(A^*) - \frac{\varkappa}{2}$$

be fulfilled, whence we obtain

$$I \geq \int_{X_2(l)} \theta\left[\nu''(A^*) - P(A^*) + \frac{\varkappa}{2}\right] dP(X_2^l)$$

$$= \sum_{k/l > P(A^*) - \varkappa/2} C_l^k [P(A^*)]^k [1 - P(A^*)]^{l-k}.$$

As is known, the last sum exceeds $\frac{1}{2}$ provided only that $l > 2/\varkappa$. Returning to (A.8), we obtain that for $l > 2/\varkappa$

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} \geq \frac{1}{2}\int_Q dP(X^l) = \tfrac{1}{2}P\{\pi^S(X^l) > \varkappa\},$$

q.e.d. □

§A4 Derivation of Sufficient Conditions

The following theorem is valid.

Theorem A.2. *The probability that for at least one event in the class S the frequency will deviate from the corresponding probability in an experiment of size l by an amount exceeding $\varkappa$ is bounded by*

$$P\{\pi^S(x_1, \ldots, x_l) > \varkappa\} < 6m^S(2l)e^{-\varkappa^2 l/4}. \tag{A.9}$$

Corollary. *In order that the frequency of events in class S shall converge (in probability) to the corresponding probabilities uniformly over the class S, it is sufficient that there exist finite n such that for $l > n$*

$$m^S(l) < 1.5\frac{l^{n-1}}{(n-1)!}.$$

PROOF. In view of the basic lemma it is sufficient to bound the quantity

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X(2l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X^{2l}).$$

Consider the mapping of the space $X(2l)$ into itself obtained by a permutation T_i of the elements of the sequence X^{2l}. In view of the symmetry of the definition of the measure, the equality

$$\int_{X(2l)} f(X^{2l})\, dP(X^{2l}) = \int_{X(2l)} f(T_i X^{2l})\, dP(X^{2l})$$

holds for any integrable function $f(X)$. Therefore

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X(2l)} \frac{\sum_{i=1}^{(2l)!} \theta\left[\rho^S(T_i X^{2l}) - \frac{\varkappa}{2}\right]}{(2l)!} dP(X^{2l}), \tag{A.10}$$

where the sum is taken over all $(2l)!$ permutations.

First we observe that

$$\theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] = \theta\left[\sup_A |\nu'(A) - \nu''(A)| - \frac{\varkappa}{2}\right]$$
$$= \sup_A \theta\left[|\nu'(A) - \nu''(A)| - \frac{\varkappa}{2}\right]$$

Clearly if two sets A_1 and A_2 induce the same subsample on the sample $x_1, \ldots, x_l, x_{l+1}, \ldots, x_{2l}$, then

$$\nu'(A_1; T_i X^{2l}) = \nu'(A_2; T_i X^{2l}),$$
$$\nu''(A_1; T_i X^{2l}) = \nu''(A_2; T_i X^{2l}),$$

and hence

$$\rho_{A_1}(T_i X^{2l}) = \rho_{A_2}(T_i X^{2l})$$

for any permutation T_i. In other words, if two events are equivalent with respect to the sample $x_1, \ldots, x_{2l}$, then deviations of frequencies for these events are the same for all permutations T_i. Therefore if from each equivalence class one chooses one set and forms a finite system S', then

$$\sup_{A \in S} \rho_A(T_i X^{2l}) = \sup_{A \in S'} \rho_A(T_i X^{2l}).$$

The number of events in the system S' is finite and is denoted by $\Delta^{S'}(x_1, \ldots, x_{2l})$. Replacing the sup operation by a summation, we obtain

$$\sup_{A \in S} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right] = \sup_{A \in S'} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right]$$
$$\leq \sum_{A \in S'} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right].$$

These relations allow us to bound the integrand in (A.10):

$$\sup_{A \in S'} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right]$$
$$= \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \sup_{A \in S'} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right] \leq \sum_{A \in S'} \frac{\sum_{i=1}^{(2l)!} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right]}{(2l)!}$$

The expression in the square brackets is the ratio of the number of orderings in the sample (of a fixed composition) such that

$$|\nu'(A) - \nu''(A)| > \frac{\varkappa}{2},$$

to the total number of permutations. It is easy to see that this expression is equal to

$$\Gamma^* = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

$$k: \left\{ \left| \frac{k}{l} - \frac{m-k}{l} \right| > \frac{\varkappa}{2} \right\},$$

where m equals the number of elements in the sample $x_1, \ldots, x_{2l}$ belonging to A.

In Section A.5 we bound the expression Γ, with the result that

$$\Gamma^* < 3 \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}.$$

Thus

$$\sum_{A \in S'} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2} \right] < \sum_{A \in S'} 3 \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}$$

$$= 3\Delta^S(x_1, \ldots, x_{2l}) \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}$$

$$\leq 3m^S(2l) \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}.$$

Substituting this bound into the integral (A.10), we obtain

$$P\left\{ \rho^S(X^{2l}) > \frac{\varkappa}{2} \right\} < 3m^S(2l) \exp\left\{ -\frac{\varkappa^2 l}{4} \right\},$$

whence in view of the basic lemma

$$P\{\pi(X^l) > \varkappa\} < 6m^S(2l) \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}.$$

The theorem is proved. □

PROOF OF THE COROLLARY. Let n exist such that for $l > n$

$$m^S(l) < 1.5 \frac{l^{n-1}}{(n-1)!}.$$

Then clearly

$$\lim_{l \to \infty} P\{\pi^S(X^l) > \varkappa\} < 9 \lim_{l \to \infty} \frac{(2l)^{n-1}}{(n-1)!} \exp\left\{ -\frac{\varkappa^2 l}{4} \right\} = 0,$$

i.e., the uniform convergence in probability is valid. □

The sufficient condition obtained does not depend on the properties of the distribution (the only condition is the measurability of functions π^S and ρ^S), but depends on the inner properties of the system S.

Remark. As it was proved in Section A.2 only if the function $m^S(l)$ is not identically 2^l, there exists n such that for $l > n$

$$m^S(l) < 1.5 \frac{l^{n-1}}{(n-1)!}.$$

Therefore the sufficient condition is always fulfilled when

$$m^S(l) \not\equiv 2^l.$$

§A5 A Bound on the Quantity Γ

We bound the expression

$$\Gamma = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

where k runs over the values satisfying the inequalities

$$\left| \frac{k}{l} - \frac{m-k}{l} \right| > \varkappa, \qquad \max(0, m-l) \le k \le \min(m, l),$$

or equivalently the inequalities

$$\left| k - \frac{m}{2} \right| > \frac{\varkappa l}{2}, \qquad \max(0, m-l) \le k \le \min(m, l),$$

and l and $m \le 2l$ are arbitrary positive integers.

We decompose Γ into two summands, $\Gamma = \Gamma_1 + \Gamma_2$.

$$\Gamma_1 = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l}, \quad \text{where } k > \frac{\varkappa l}{2} + \frac{m}{2}.$$

$$\Gamma_2 = \sum_k^k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l}, \quad \text{where } k < \frac{\varkappa l}{2} - \frac{m}{2}.$$

Introduce the notation

$$p(k) = \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l} \tag{A.11}$$

$$q(k) = \frac{p(k+1)}{p(k)} = \frac{(m-k)(l-k)}{(k+1)(l+k+1-m)}, \tag{A.12}$$

where

$$\max(0, m-l) \le k \le \min(m, l).$$

Furthermore denote

$$s = \min(m, l), \qquad T = \max(0, m - l);$$

$$d(k) = \sum_{i=k}^{s} p(i).$$

Clearly the relation

$$d(k+1) = \sum_{i=k+1}^{s} p(i) = \sum_{i=k}^{s-1} p(i+1) = \sum_{i=k}^{s-1} p(i)q(i) \tag{A.13}$$

is valid. Furthermore it follows directly from (A.12) that for $i < j$, $q(i) < q(j)$, i.e., $q(i)$ is monotonically decreasing. Therefore the inequality

$$d(k+1) = \sum_{i=k}^{s-1} p(i)q(i) < q(k) \sum_{i=k}^{s} p(i)$$

follows from (A.13). Furthermore by definition of $d(k)$ we have

$$d(k+1) < q(k)\, d(k).$$

Applying this relation successively, we obtain for arbitrary k and j satisfying the condition $T \leq j < k \leq s$

$$d(k) < d(j) \prod_{i=j}^{k-1} q(i).$$

Furthermore, since $d(j) \leq 1$,

$$d(k) < \prod_{i=j}^{k-1} q(i), \tag{A.14}$$

where j is an arbitrary integer smaller than k.

Set

$$t = k - \frac{m-1}{2}.$$

Then

$$q(t) = \frac{\frac{m+1}{2} - t}{\frac{m+1}{2} + t} \cdot \frac{\left(l - \frac{m-1}{2}\right) - t}{\left(l - \frac{m-1}{2}\right) + t}.$$

Moreover, as long as $T < k < s$, the inequality

$$|t| < \min\left(\frac{m+1}{2}, l - \frac{m-1}{2}\right)$$

is clearly valid.

To approximate $q(k)$ we study the function

$$F(t) = \frac{a-t}{a+t} \cdot \frac{b-t}{b+t},$$

assuming that a and b are both positive.

For $|t| < \min(a, b)$

$$\ln F(t) = \ln(a-t) - \ln(a+t) + \ln(b-t) - \ln(b+t).$$

Furthermore we have

$$\ln F(0) = 0, \qquad \frac{d}{dt}(\ln F(t)) = -\left[\frac{2a}{a^2-t^2} + \frac{2b}{b^2-t^2}\right].$$

This implies that for $|t| < \min(a, b)$

$$\frac{d}{dt}(\ln F(t)) \leq -2\left[\frac{1}{a} + \frac{1}{b}\right].$$

Correspondingly for $|t| < \min(a, b)$ and $t \geq 0$ the inequality

$$\ln F(t) \leq -2\left[\frac{1}{a} + \frac{1}{b}\right]t$$

is fulfilled.

Returning to $q(t)$, we obtain for $t \geq 0$

$$\ln q(t) \leq -2\left[\frac{2}{m+1} + \frac{2}{2l-m+1}\right]t = -8\,\frac{l+1}{(m+1)(2l-m+1)}\,t.$$

We now bound

$$\ln\left(\prod_{i=j}^{k-1} q(i)\right),$$

assuming that $(m-1)/2 \leq j \leq k-1$:

$$\ln\left(\prod_{i=j}^{k-1} q(i)\right) = \sum_{i=j}^{k-1} \ln q(i)$$

$$\leq \frac{-8(l+1)}{(m+1)(2l-m+1)} \sum_{i=j}^{k-1} \left(i - \frac{m-1}{2}\right).$$

Returning to (A.14), we obtain

$$\ln d(k) < \frac{-8(l+1)}{(m+1)(2l-m+1)} \sum_{i=j}^{k-1} \left(i - \frac{m-1}{2}\right);$$

here j is an arbitrary number smaller than k. Therefore for $k > (m-1)/2$ one can set $j = (m-1)/2$ for m odd and $j = m/2$ for m even, obtaining a

stronger bound. Next, summing the arithmetic progression, we obtain

$$\ln d(k) < \begin{cases} -\dfrac{4(l+1)}{(m+1)(2l-m+1)}\left(k-\dfrac{m}{2}+1\right)^2 & \text{for even } m, \\ -\dfrac{4(l+1)}{(m+1)(2l-m+1)}\left(k-\dfrac{m-1}{2}+1\right)\left(k-\dfrac{m-1}{2}\right) & \text{for odd } m. \end{cases}$$

Finally Γ_1 is $d(k)$ for the first integer k such that

$$k-\frac{m}{2} > \frac{\varkappa^2 l}{2},$$

whence

$$\ln \Gamma_1 < -\frac{l+1}{(m+1)(2l-m+1)}\varkappa^2 l^2.$$

In the same manner one can bound Γ_2, since the distribution (A.11) is symmetric with respect to the point $k = m/2$. Thus

$$\Gamma < 2\exp\left\{-\frac{(l+1)\varkappa^2 l^2}{(m+1)(2l-m+1)}\right\}. \tag{A.15}$$

The right-hand side of (A.15) attains its maximum at $m = l$, and consequently

$$\Gamma < 2\exp\left\{-\frac{\varkappa^2 l^2}{l+1}\right\} < 3\exp\{-\varkappa^2 l\}.$$

§A6 A Bound on the Probability of Uniform Relative Deviation

In this section we shall prove

Theorem A.3. *For any p $(1 < p \leq 2)$ the bound*

$$P\left\{\sup_{A\in S}\frac{P(A)-\nu(A)}{\sqrt[p]{P(A)}} > \varkappa\right\} < 8m^S(2l)\exp\left\{-\frac{\varkappa^2}{4}l^{2-(2/p)}\right\} \tag{A.16}$$

is valid.

PROOF. Consider two events constructed from a random and independent sample of size $2l$: The event Q_1:

$$Q_1 = \left\{\sup_{A\in S}\frac{P(A)-\nu'(A)}{\sqrt[p]{P(A)}} > \varkappa\right\}$$

and the event Q_2:

$$Q_2 = \left\{ \sup_{A \in S} \frac{|\nu'(A) - \nu''(A)|}{\sqrt[p]{\nu(A) + 1/2l}} > \varkappa \right\},$$

where $\nu'(A)$ is the frequency of the event A computed from the first half-sample of length l; $\nu''(A)$ is the frequency of the event A computed from the second half-sample; $\nu''(A)$ is the frequency of the event computed from the sample of length $2l$.

Observe that in the case $l \leq \varkappa^{-p/(p-1)}$ the theorem is trivial. Accordingly we shall prove the theorem as follows: First we show that for $l > \varkappa^{-p/(p-1)}$ the inequality

$$P(Q_1) < 4P(Q_2)$$

is valid, and then we bound the probability of the event Q_2. Thus we shall prove the lemma:

Lemma A.2. *For $l > \varkappa^{-p/(p-1)}$ the inequality*

$$P(Q_1) < 4P(Q_2) \tag{A.17}$$

is valid.

PROOF. Assume that event Q_1 occurred. This means that there exists A^* such that for the first half sample the inequality

$$P(A^*) - \nu'(A^*) > \varkappa \sqrt[p]{P(A^*)}$$

is fulfilled. Since $\nu'(A) \geq 0$, this implies that

$$P(A^*) > \varkappa^{p/(p-1)}.$$

Assume that for the second half sample the frequency of occurrence of event A^* exceeds the probability $P(A^*)$:

$$\nu''(A^*) > P(A^*).$$

Recall now that $l > \varkappa^{-p/(p-1)}$. Under these conditions event Q_2 will definitely occur.

To show this we bound the quantity

$$\mu = \frac{|\nu'(A^*) - \nu''(A^*)|}{\sqrt[p]{\nu(A^*) + 1/2l}} < \frac{\nu''(A^*) - \nu'(A^*)}{\sqrt[p]{\nu(A^*) + 1/2l}} \tag{A.18}$$

under the conditions

$$\nu'(A^*) < P(A^*) - \varkappa \sqrt[p]{P(A^*)}$$

$$\nu''(A^*) > P(A^*),$$

$$P(A^*) > \varkappa^{p/(p-1)}.$$

For this purpose we find the minimum of the function

$$T = \frac{x - y}{\sqrt[p]{x + y + c}}$$

in the domain $0 < a \leq x \leq 1, 0 < y \leq b, c > 0$. We have for $p > 1$

$$\frac{\partial T}{\partial x} = \frac{1}{p} \frac{(p-1)x + (p+1)y + pc}{(x + y + c)^{(p+1)/p}} > 0,$$

$$\frac{\partial T}{\partial y} = -\frac{1}{p} \frac{(p+1)x + (p+1)y + pc}{(x + y + c)^{(p+1)/p}} < 0.$$

Consequently T attains its minimum in the admissible domain for $x = a$ and $y = b$. Therefore the quantity μ will be bounded from below if one replaces $\nu'(A^*)$ by $P(A^*) - \varkappa\sqrt[p]{P(A^*)}$ and $\nu''(A^*)$ by $P(A^*)$ in (A.18). Thus

$$\mu > \frac{\varkappa\sqrt[p]{2P(A^*)}}{\sqrt[p]{2P(A^*) - \varkappa\sqrt[p]{P(A^*)} + 1/l}}.$$

Furthermore, since $P(A^*) > \varkappa^{p/(p-1)}$, $l > \varkappa^{-p/(p-1)}$, we have

$$\mu > \frac{\varkappa\sqrt[p]{2P(A^*)}}{\sqrt[p]{2P(A^*) - \varkappa^{p/(p-1)} + \varkappa^{p/(p-1)}}} = \varkappa.$$

Thus if Q_1 occurs and the conditions $P(A^*) \leq \nu''(A^*)$ and $l > \varkappa^{-p/(p-1)}$ are fulfilled, then Q_2 occurs as well.

Observe that the second half sample is chosen independently of the first and, as is known, for $l > 2/P(A^*)$ the frequency of occurrence of the event A^* exceeds $P(A^*)$ with probability $\frac{1}{4}$. Therefore, provided Q_1 is fulfilled, the event

$$\nu''(A^*) > P(A^*)$$

occurs with probability exceeding $\frac{1}{4}$ as long as $l > \varkappa^{-p/(p-1)}$. Thus for $l > \varkappa^{p/(p-1)}$

$$P(Q_2) > \tfrac{1}{4}P(Q_1).$$

The lemma is proved. □

Lemma A.3. *For any p $(1 < p \leq 2)$ the bound*

$$P(Q_2) < 2m^S(2l) \exp\left\{-\frac{\varkappa^2}{4} l^{2-(2/p)}\right\}$$

is valid.

PROOF. Denote by $R_A(X^{2l})$ the quantity

$$R_A(X^{2l}) = \frac{|\nu'(A) - \nu''(A)|}{\sqrt[p]{\nu(A) + 1/2l}}.$$

Then the estimated probability equals

$$P(Q_2) = \int_{X(2l)} \theta\left[\sup_{A \in S} R_A(X^{2l}) - \varkappa\right] dP(X^{2l}).$$

Here the integration is carried out over the space of all possible samples of size $2l$.

Consider now all possible permutations T_i ($i = 1, 2, \ldots, (2l)!$) of the sequence $x_1, \ldots, x_{2l}$. For each such permutation T_i the equality

$$\int_{X(2l)} \theta\left[\sup_{A \in S} R_A(X^{2l}) - \varkappa\right] dP(X^{2l}) = \int_{X(2l)} \theta\left[\sup_{A \in S} R_A(T_i X^{2l}) - \varkappa\right] dP(X^{2l})$$

is valid. Therefore the equality

$$\int_{X(2l)} \theta\left[\sup_{A \in S} R_A(X^{2l}) - \varkappa\right] dP(X^{2l})$$

$$= \int_{X(2l)} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left[\sup_{A \in S} R_A(T_i X^{2l}) - \varkappa\right] dP(X^{2l})$$

is valid.

Consider now the integrand. Since the sample $x_1, \ldots, x_{2l}$ is fixed, instead of the system of events S one can consider a finite system of events S' which contains one representative for each one of the equivalence classes. Thus the equality

$$\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left[\sup_{A \in S} R_A(T_i X^{2l}) - \varkappa\right] = \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left[\sup_{A \in S'} R_A(T_i X^{2l}) - \varkappa\right]$$

is valid. Furthermore

$$\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left[\sup_{A \in S'} R_A(T_i X)^{2l} - \varkappa\right] < \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \sum_{A \in S'} \theta[R_A(T_i X^{2l}) - \varkappa]$$

$$= \sum_{A \in S'} \left\{\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta[R_A(T_i X^{2l}) - \varkappa]\right\}. \tag{A.19}$$

The expression in the braces is the probability of the deviation of frequencies in two half samples for a fixed event A and a given composition of the complete sample. This probability equals

$$\Gamma = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

where m is the number of occurrences of events A in the complete sample, and k is the number of occurrences of the events in the first half sample; k runs

over the values

$$\max(0, m - l) \leq k \leq \min(m, l),$$

$$\frac{\left|\frac{k}{l} - \frac{m-k}{l}\right|}{\sqrt[p]{\frac{m+1}{2l}}} > \varkappa.$$

Denote by $\varkappa'$ the quantity

$$\sqrt[p]{\frac{m+1}{2l}}\,\varkappa = \varkappa'.$$

Using this notation the restrictions become

$$\max(0, m - l) \leq k \leq \min(m, l),$$

$$\left|\frac{k}{l} - \frac{m-k}{l}\right| > \varkappa'. \tag{A.20}$$

In Section A.5 the following bound on the quantity Γ under the restrictions (A.20) was obtained:

$$\Gamma < 2\exp\left\{-\frac{(1+1)(\varkappa')^2 l^2}{(m+1)(2l-m+1)}\right\}. \tag{A.21}$$

Expressing (A.19) in terms of $\varkappa$, we obtain

$$\Gamma < 2\exp\left\{-\frac{\varkappa^2(l+1)l^2}{2(2l-m+1)(m+1)}\left(\frac{m+1}{2l}\right)^{2/p}\right\}.$$

The right-hand side of the inequality attains its maximum at $m = 0$. Thus

$$\Gamma < 2\exp\left\{-\frac{\varkappa^2}{4} l^{2-(2/p)}\right\}. \tag{A.22}$$

Substituting (A.22) into the right-hand side of (A.19) and integrating, we have

$$P(Q_2) < 2m^s(2l)\exp\left\{-\frac{\varkappa^2}{4} l^{2-(2/p)}\right\}. \tag{A.23}$$

The lemma is thus proved. □

The inequalities (A.17) and (A.23) yield the assertion of the theorem. □

Chapter 7

A Method of Minimizing Empirical Risk for the Problem of Regression Estimation

§1 Uniform Convergence of Means to Mathematical Expectations

In this book the problem of pattern recognition is formulated as the simplest problem of estimating dependences from empirical data. The simplicity of the problem is due to the fact that it reduces to minimizing the functional

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy, \tag{7.1}$$

with an unknown density $P(x, y)$, from the sample

$$x_1, y_1; \ldots; x_l, y_l, \tag{7.2}$$

when y takes on only two values 0 and 1 and $F(x, \alpha)$ is a class of indicator functions.

The problem of regression estimation is considered to be more complex. It also reduces to minimizing a functional with unknown density $P(x, y)$ on the basis of the sample (7.2), but in this case y may take on an arbitrary value and the class $F(x, \alpha)$ consists of square-integrable functions. Therefore the construction of the theory of minimizing the risk (7.1) in a class of not necessarily indicator functions $F(x, \alpha)$ by means of minimization of an empirical functional

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2 \tag{7.3}$$

can be viewed as a generalization of results of the theory obtained in the preceding chapter to a wider class of functions. In this chapter we shall

construct the theory of regression estimation using the method of minimizing the empirical risk (7.2) as a natural generalization of the solution for the pattern recognition problem.

This is our first opportunity to implement this approach. It was not possible to do this utilizing parametric methods as in problems of pattern recognition (Chapter 3) and regression estimation (Chapters 4 and 5). Solutions of problems were carried out there under stipulations of intrinsically different models for densities $P(x, y)$: in the pattern recognition problem the structure of the density was determined by a union of two densities; in the regression estimation problem it was given by a measurement model with additive noise. Here, however, the principle for solving the problem is the same: a search for a function which minimizes (7.1) is carried out by means of minimizing the empirical functional (7.3).

In the preceding chapter conditions were obtained under which this approach can be successfully implemented for a class of indicator functions $F(x, \alpha)$. Now we shall obtain conditions which assure a successful application of the method of minimizing empirical risk when the class $F(x, \alpha)$ is of a more general nature.

In the problem of pattern recognition, the functional (7.1) determines for each fixed α the probability of a certain event (an incorrect classification of the vector which is to be "recognized"), and the empirical functional (7.3) determines the frequency of this event computed from the sample. Conditions for applicability of the method of minimizing empirical risk are associated here with the uniform convergence, over a class of events, of frequencies of events to their probabilities.

In the problem of regression estimation the functional (7.1) determines for each fixed α the mathematical expectation of the random variable

$$\xi(\alpha) = (y - F(x, \alpha))^2,$$

and the empirical functional (7.3) determines the empirical mean of this random variable computed from the sample (7.2).

Above (Chapter 6, Section 1) it was shown that a successful application of the method of minimizing an empirical risk might be associated with the validity of the uniform convergence of the means to their mathematical expectations:

$$P\left\{\sup_{\alpha} |I(\alpha) - I_{\text{emp}}(\alpha)| > \varkappa\right\} < \eta(l, \varkappa),$$
$$\lim_{l \to \infty} \eta(l, \varkappa) = 0. \tag{7.4}$$

It was shown that under the condition (7.4) the value of the functional (7.1) at the point of empirical minimum $F(x, \alpha_{\text{emp}})$ deviates with probability $1 - \eta$ from the minimal value of $I(\alpha_0)$ in the class $F(x, \alpha)$ by an amount not exceeding $2\varkappa$:

$$P\{I(\alpha_{\text{emp}}) - I(\alpha_0) > 2\varkappa\} < \eta.$$

Thus the problem is reduced to the determination of the conditions for the existence of uniform convergence of the means to their mathematical expectations and to the estimation of the rate of convergence.

As in the previous chapter the validity of basic theorems on uniform convergence does not depend on the form of the loss function. Therefore, in spite of a quadratic loss function used in the text a general theory is obtained.

§2 A Particular Case

As above, we shall start with simple case: the set of functions $F(x, \alpha)$ consists of a finite number N of elements

$$F(x, \alpha_1), \ldots, F(x, \alpha_N).$$

For this case the inequality

$$P\left\{\sup_i |I(\alpha_i) - I_{\text{emp}}(\alpha_i)| > \varkappa\right\} < \sum_{i=1}^{N} P\{|I(\alpha_i) - I_{\text{emp}}(\alpha_i)| > \varkappa\}$$
$$\leq N \sup_i P\{|I(\alpha_i) - I_{\text{emp}}(\alpha_i)| > \varkappa\} \quad (7.5)$$

is valid.

In Chapter 6, for an analogous situation of bounding the rate of uniform convergence of frequencies of events to their probabilities, a nontrivial bound on the second factor was used. In this case a nontrivial bound on

$$\sup_i P\{|I(\alpha_i) - I_{\text{emp}}(\alpha_i)| > \varkappa\}$$

is generally unavailable—since the random variable $I_{\text{emp}}(\alpha_i)$ may possess "large deviations", and therefore its deviation from the mean $I(\alpha_i)$ may be arbitrary. We have already encountered such a situation in Chapter 2, where it was necessary to take into account the measure of "possible large deviations" when determining a guaranteed bound on the mathematical expectation based on the value of the empirical mean. In particular it was shown (cf. Chapter 2, Section 2) that for this purpose it is sufficient to know either a bound on possible losses,

$$\sup_{\alpha, x, y} (y - F(x, \alpha))^2 \leq \tau,$$

or a bound on the relative variance of losses,

$$\sup_\alpha \sqrt{\frac{\int (y - F(x, \alpha))^4 P(x, y)\, dx\, dy}{(\int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy)^2}} - 1 \leq \tau.$$

Thus to obtain a bound on the rate of uniform convergence of the means to their mathematical expectations the prior information on the magnitude of

possible large deviations should be utilized. We remark that for solving the problem of pattern recognition there was no need for such information. In view of the statement of the problem, the loss function $(y - F(x, \alpha))^2$ was bounded by 1, i.e., the prior information about the large deviations was contained in the statement of the problem.

In this chapter we shall utilize both types of prior information on large deviations, and for each of them obtain a bound on the rate of uniform convergence.

The simplest condition under which it is possible to obtain a bound on the rate of uniform convergence of the means to mathematical expectations is the condition of uniform boundedness of the losses.†

$$(y - F(x, \alpha))^2 \leq \tau \tag{7.6}$$

for all α, $x \in X$ and $y \in Y$.

Let the inequality (7.6) hold. We show that in this case the bound

$$P\left\{\sup_i |I(\alpha_i) - I_{\text{emp}}(\alpha_i)| > \varkappa\tau\right\} < 18Nle^{-\varkappa^2 l/4}$$

is valid. To obtain this bound we write the functionals $I(\alpha_i)$ and $I_{\text{emp}}(\alpha_i)$ using the Lebesgue integrals:

$$\begin{aligned} I(\alpha_i) &= \lim_{n\to\infty} \sum_{j=1}^{n} \frac{\tau}{n} P\left\{(y - F(x, \alpha_i))^2 > \frac{j\tau}{n}\right\}, \\ I_{\text{emp}}(\alpha_i) &= \lim_{n\to\infty} \sum_{j=1}^{n} \frac{\tau}{n} \nu\left\{(y - F(x, \alpha_i))^2 > \frac{j\tau}{n}\right\}, \end{aligned} \tag{7.7}$$

where $\nu\{(y - F(x, a_i))^2 > j\tau/n\}$ denotes the frequency of the event $\{(y - F(x, \alpha_i))^2 > j\tau/n\}$ computed from the sample (7.2). Denote by $A_{\alpha_i, j}$ the event

$$\left\{(y - F(x, \alpha_i))^2 > \frac{j\tau}{n}\right\}.$$

Then in view of (7.7)

$$\begin{aligned} |I(\alpha_i) - I_{\text{emp}}(\alpha_i)| &\leq \lim_{n\to\infty} \sum_{j=1}^{n} \frac{\tau}{n} |P(A_{\alpha_i, j}) - \nu(A_{\alpha_i, j})| \\ &\leq \tau \sup_j |P(A_{\alpha_i, j}) - \nu(A_{\alpha_i, j})|. \end{aligned}$$

Thus

$$P\{|I(\alpha_i) - I_{\text{emp}}(\alpha_i)| > \tau\varkappa\} \leq P\left\{\sup_j |P(A_{\alpha_i, j}) - \nu(A_{\alpha_i, j})| > \varkappa\right\}.$$

† Below, various sufficient conditions for uniform convergence will be presented. Necessary and sufficient conditions are given in the appendix to this chapter.

Consider now the class of events $A_{\alpha_i,\beta}$:

$$\{(y - F(x, \alpha_i))^2 > \beta\},$$

where β is a nonnegative quantity. Clearly this class contains the events $\{A_{\alpha_i,j}\}$ whence

$$P\left\{\sup_j |P(A_{\alpha_i,j}) - \nu(A_{\alpha_i,j})| > \varkappa\right\} \leq P\left\{\sup_\beta P|(A_{\alpha_i,\beta}) - \nu(A_{\alpha_i,\beta})| > \varkappa\right\}.$$

The problem has thus been reduced to bounding the uniform convergence of frequencies to their probabilities over the class S_β of events $A_{\alpha_i,\beta}$ (with fixed values of α_i).

Utilizing the results of the preceding chapter, we bound the rate of uniform convergence of frequencies to probabilities over the class of events

$$S_\beta = \{x, y: (y - F(x, \alpha_i))^2 > \beta\}.$$

For this purpose we bound the growth function $m^{S_\beta}(l)$. Since using the rules

$$\theta[(y - F(x, \alpha_i))^2 - \beta]$$

(α_i is fixed) one can subdivide only one point x, y in all possible ways, we have in view of Theorem 6.6

$$m^{S_\beta}(l) < 1.5l.$$

Consequently, utilizing Theorem A.2 of the Appendix to Chapter 6, we obtain

$$\begin{aligned} &P\{|I(\alpha_i) - I_{\text{emp}}(\alpha_i)| > \tau\varkappa\} \\ &\quad \leq P\left\{\sup_\beta |P(A_{\alpha_i,\beta}) - \nu(A_{\alpha_i,\beta})| > \varkappa\right\} \\ &\quad < 6m^{S_\beta}(2l)e^{-\varkappa^2 l/4} < 18le^{-\varkappa^2 l/4}. \end{aligned} \tag{7.8}$$

The right-hand side of the inequality does not depend on α. Therefore, along with (7.8), a more refined bound,

$$\sup_\alpha P\{|(\alpha) - I_{\text{emp}}(\alpha)| > \tau\varkappa\} < 18le^{-\varkappa^2 l/4},$$

is valid. Returning to the bound (7.5), we have

$$P\left\{\sup_\alpha |I(\alpha) - I_{\text{emp}}(\alpha)| > \tau\varkappa\right\} < 18Nle^{-\varkappa^2 l/4}.$$

We shall require that this probability be equal to η:

$$18Nle^{-\varkappa^2 l/4} = \eta.$$

Therefore the deviation $\varkappa$ should not be less than

$$\varkappa = 2\sqrt{\frac{\ln N + \ln l - \ln(\eta/18)}{l}}.$$

The result obtained can be stated as

Theorem 7.1. *Let the class $F(x, \alpha)$ consist of N functions for which the losses $(y - F(x, \alpha))^2$ in the domain $x \in X$, $y \in Y$ are uniformly bounded by a constant τ. Then one can assert with probability $1 - \eta$ that the inequality*

$$I_{\text{emp}}(\alpha_i) - 2\tau\sqrt{\frac{\ln N + \ln l - \ln(\eta/18)}{l}} < I(\alpha_i) < I_{\text{emp}}(\alpha_i) + 2\tau\sqrt{\frac{\ln N + \ln l - \ln(\eta/18)}{l}}$$

is valid simultaneously for all N functions $F(x, \alpha_i)$.

Remark. The theorem is valid simultaneously for all N functions, including the function $F(x, \alpha_{\text{emp}})$ which yields the minimum for the value of the empirical risk. Hence the inequality

$$I(\alpha_{\text{emp}}) < I_{\text{emp}}(\alpha_{\text{emp}}) + 2\tau\sqrt{\frac{\ln N + \ln l - \ln(\eta/18)}{l}}$$

is valid. Thus if the loss function is uniformly bounded and the number of functions $F(x, \alpha_i)$ in the class is finite, then the uniform convergence of the means to their mathematical expectations holds. Theorem 7.1 is a direct generalization of Theorem 6.1.

§3 A Generalization to a Class with Infinitely Many Members

Now let the class $F(x, \alpha)$ consist of infinitely many elements while admitting a cover by a finite ε-net in either the C metric or the L_p^2 metric. As before, let the restriction (7.6) be valid. We show that in this case a bound on the quality of the rule minimizing the empirical risk exists which is analogous to the one that follows from Theorem 7.1.

Theorem 7.2. *Let the set of functions $F(x, \alpha)$ be covered by a finite ε-net $F(x, \alpha_1), \ldots, F(x, \alpha_{N(\varepsilon)})$. Then with probability $1 - \eta$ the quality of the function*

$F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk is bounded by

$$I(\alpha_{\text{emp}}) < I_{\text{emp}}(\alpha_i(\alpha_{\text{emp}})) + 2\tau\sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln(\eta/18)}{l}} + 2\varepsilon\sqrt{\tau},$$

where $F(x, \alpha_i(\alpha_{\text{emp}}))$ is a function in the ε-net closest to $F(x, \alpha_{\text{emp}})$.

The *proof* is carried out along the lines of the proof of Theorem 6.4.

(1) Select on the set of functions $F(x, \alpha)$ a finite ε-net consisting of $N(\varepsilon)$ elements

$$F(x, a_1), \ldots, F(x, \alpha_{N(\varepsilon)}).$$

According to Theorem 7.1 the inequalities

$$I(\alpha_i) < I_{\text{emp}}(\alpha_i) + 2\tau\sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln(\eta/18)}{l}} \tag{7.9}$$

are valid simultaneously for all elements of the ε-net with probability $1 - \eta$.

(2) We bound the amount of deviation of the functionals $I(\alpha_1)$ and $I(\alpha_2)$ for functions $F(x, \alpha_1)$ and $F(x, \alpha_2)$ separated from each other by at most ε, i.e., we find the smallest $\delta(\varepsilon)$ such that the inequality

$$|I(\alpha_1) - I(\alpha_2)| \leq \delta(\varepsilon)$$

is fulfilled provided only the conditions

$$\rho_L(\alpha_1, \alpha_2) = \left(\int (F(x, \alpha_1) - F(x, a_2))^2 P(x)\, dx\right)^{1/2} \leq \varepsilon \tag{7.10}$$

$$\left(\rho_C(\alpha_1, \alpha_2) = \sup_x |F(x_1\alpha_1) - F(x_1, \alpha_2)| \leq \varepsilon\right)$$

are satisfied. For this purpose we carry out the transformations

$$\begin{aligned} |I(\alpha_1) - I(\alpha_2)| &= \left| \int (y - F(x, \alpha_1))^2 P(x, y)\, dx\, dy \right. \\ &\quad \left. - \int (y - F(x, \alpha_2))^2 P(x, y)\, dx\, dy \right| \\ &= \left| \int (F(x, \alpha_1) - F(x, \alpha_2)) \right. \\ &\quad \left. \times (2y - F(x, \alpha_1) - F(x, \alpha_2)) P(x, y)\, dx\, dy \right| \\ &\leq \varepsilon \sqrt{\int (2y - F(x, \alpha_1) - F(x, \alpha_2))^2 P(x, y)\, dx\, dy}. \end{aligned}$$

Here we have utilized the Cauchy–Schwarz inequality and the bound (7.10). Next we utilize the convexity of the function $(y - F(x, \alpha))^2$:

$$\int (2y - F(x, \alpha_1) - F(x, \alpha_2))^2 P(x, y)\, dx\, dy$$

$$\leq 2 \int (y - F(x, \alpha_1))^2 P(x, y)\, dx\, dy$$

$$+ 2 \int (y - F(x, \alpha_2))^2 P(x, y)\, dx\, dy.$$

We thus obtain

$$|I(\alpha_1) - I(\alpha_2)| \leq \varepsilon \sqrt{2(I(\alpha_1) + I(\alpha_2))}. \tag{7.11}$$

However, by the condition, $I(\alpha) \leq \tau$. Finally we obtain

$$|I(\alpha_1) - I(\alpha_2)| \leq 2\varepsilon \sqrt{\tau}. \tag{7.11a}$$

(3) Now let $F(x, \alpha_{\text{emp}})$ be the function which yields the minimum for the empirical risk. We choose a function $F(x, \alpha_i(\alpha_{\text{emp}}))$ in the ε-net $F(x, \alpha_1), \ldots, F(x, \alpha_{N(\varepsilon)})$ closest to $F(x, \alpha_{\text{emp}})$. For this function the inequality (7.9) is satisfied with probability $1 - \eta$. We strengthen this inequality utilizing the bound (7.11a). This leads to

$$I(\alpha_{\text{emp}}) < I_{\text{emp}}(\alpha_i(\alpha_{\text{emp}})) + 2\varepsilon \sqrt{\tau} + 2\tau \sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln(\eta/18)}{l}}. \tag{7.12}$$

The theorem is proved. □

Remarks. The theorem is valid for any ε (assigned before sampling). Therefore ε may be selected from the condition of the minimum for the expression

$$r(\varepsilon) = \varepsilon \sqrt{\tau} + \tau \sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln(\eta/18)}{l}}.$$

Note also that for any set $F(x, \alpha)$ and any ε the minimal number of elements in an ε-net constructed in the L_p^2 metric does not exceed the minimal number of elements in an ε-net in the C metric. Therefore the bound (7.12) is more precise if the ε-net is constructed in the L_p^2 metric. However, in order to define this metric the density $P(x)$ should be known.

§4 The Capacity of a Set of Arbitrary Functions

In Chapter 6 we introduced the notion of *capacity* for a set of indicator functions. The capacity was determined by a maximal number of points $x_1, \ldots, x_h$ which can be subdivided in all possible ways into two classes by means of a given set of indicator functions.

We shall now extend the notion of capacity to sets of functions $F(x, \alpha)$ of an arbitrary nature. For this purpose we shall introduce the following parametric set of indicator functions:

$$\hat{F}(x, y; \alpha, \beta) = \theta((y - F(x, \alpha))^2 + \beta)$$

in the parameters α and β (β is a real number).

Definition. The capacity of the set of indicator functions $\hat{F}(x, y; \alpha, \beta)$ is called the capacity of the set $F(x, \alpha)$.

Thus the capacity of the set $F(x, \alpha)$ determines the largest number h of pairs x_i, y_i which can be subdivided in all possible ways into two classes by means of the rules $\hat{F}(x, y; \alpha, \beta)$.

The capacity of a set of functions linear in its parameters,

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x),$$

equal $n + 1$.

Under this definition of capacity, the growth function for the system of events

$$S_{\alpha, \beta} = \{x, y: (y - F(x, \alpha))^2 > \beta\}$$

is bounded by

$$m^{S_{\alpha, \beta}}(l) < 1.5 \frac{l^h}{h!}$$

for $l > h$. Let the capacity of a set of functions $F(x, \alpha)$ be equal to h, and as above, let the loss function be bounded by τ. Under these conditions the following theorem is valid.

Theorem 7.3. *For $l > h$ simultaneously for the whole class of functions $F(x, \alpha)$, the inequality*

$$I_{\text{emp}}(\alpha) - 2\tau\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}} < I(\alpha) < I_{\text{emp}}(\alpha) + 2\tau\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}}$$

is satisfied with probability $1 - \eta$.

PROOF: We express functionals $I(\alpha)$ and $I_{\text{emp}}(\alpha)$ in terms of Lebesgue integrals:

$$I(\alpha) = \lim_{n \to \infty} \sum_{i=1}^{n} \frac{\tau}{n} P\left\{(y - F(x, \alpha))^2 > \frac{i\tau}{n}\right\},$$

$$I_{\text{emp}}(\alpha) = \lim_{n \to \infty} \sum_{i=1}^{n} \frac{\tau}{n} \nu\left\{(y - F(x, \alpha))^2 > \frac{i\tau}{n}\right\}.$$

Here $P\{(y - F(x, \alpha))^2 > i\tau/n\}$ denotes the probability of the event $\{(y - F(x, \alpha))^2 > i\tau/n\}$, and $\nu\{(y - F(x, \alpha))^2 > i\tau/n\}$ is the frequency of this event computed from the training sequence.

The event

$$\{(y - F(x, \alpha))^2 > \beta\}$$

will be denoted by $A_{\alpha,\beta} \in S_{\alpha\beta}$. Then

$$|I(\alpha) - I_{\text{emp}}(\alpha)| \leq \lim_{n\to\infty} \sum_{i=1}^{n} \frac{\tau}{n} |P(A_{\alpha,i}) - \nu(A_{\alpha,i})|.$$

Whence

$$|I(\alpha) - I_{\text{emp}}(\alpha)| \leq \tau \sup_{\beta} |P(A_{\alpha,\beta}) - \nu(A_{\alpha,\beta})|.$$

Furthermore it follows that

$$P\left\{\sup_{\alpha} |I(\alpha) - I_{\text{emp}}(\alpha)| > \tau\varkappa\right\}$$

$$\leq P\left\{\sup_{\alpha,\beta} |P(A_{\alpha,\beta}) - \nu(A_{\alpha,\beta})| > \varkappa\right\}.$$

Since for $l > h$ the growth function of the system of events $S_{\alpha,\beta}$ is bounded by $1.5l^h/h!$, utilizing Theorem A.2 of the Appendix to Chapter 6 we obtain

$$P\left\{\sup_{\alpha} |I(\alpha) - I_{\text{emp}}(\alpha)| > \tau\varkappa\right\}$$

$$< 6m^S(2l)e^{-\varkappa^2 l/4} < 9\frac{(2l)^h}{h!}e^{-\varkappa^2 l/4}. \tag{7.13}$$

Setting the right-hand side of the inequality equal to η and solving the resulting equation for $\varkappa$, we have

$$\varkappa = 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}}. \tag{7.14}$$

It thus follows from (7.13) and (7.14) that for $l > h$ the inequality

$$I_{\text{emp}}(\alpha) - 2\tau\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}} < I(\alpha)$$

$$< I_{\text{emp}}(\alpha) + 2\tau\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}}$$

is satisfied with probability $1 - \eta$ simultaneously for all functions of the set $F(x, \alpha)$. The theorem is proved. □

§5 Uniform Boundedness of a Ratio of Moments

Now let for some $p > 1$ the conditions

$$\sup_{\alpha} \frac{\sqrt[p]{\int (y - F(x, \alpha))^{2p} P(x, y)\, dx\, dy}}{\int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy} \leq \tau \tag{7.15}$$

be fulfilled, i.e., for any fixed $\alpha = \alpha^*$ let the ratio of the pth order mean† of the random variable

$$\xi(\alpha^*) = (y - F(x, \alpha^*))^2$$

to the first order mean be bounded by τ. The fulfillment of the conditions (7.15) is the basic requirement imposed for solving problems of dependence estimation and ill-posed problems.

In the next sections we shall show that if (7.15) holds for a $p > 1$ a theory of uniform relative deviation of the means from their mathematical expectations can be constructed. The case (7.15) for $p \geq 2$ will be the most important. For $p > 2$ maximum rate of convergence is achieved in the order of magnitude. For $p = 2$ the requirement (7.15) is equivalent to the condition of uniform boundedness of the relative variance considered in Section 2 of Chapter 2; moreover the number τ_{rel} which bounds the relative variance is related to τ, which bounds the mean of the second order, as follows:

$$\tau = \sqrt{\tau_{\text{rel}}^2 + 1}.$$

The condition (7.15) is quite weak. All parametric models of regression extimation considered in Chapters 4 and 5 satisfy this condition with τ within the narrow limits $1.35 < \tau < 2.45$ (cf. Chapter 2, Section 3).

We shall show below that if along with (7.15) one of the following three conditions is fulfilled:

(1) the set $F(x, \alpha)$ consists of a finite number of elements,
(2) the set $F(x, \alpha)$ may be covered by a finite ε-net,
(3) the set of functions $F(x, \alpha)$ possess a finite capacity,

then the method of minimizing empirical risk yields a solution to the problem of estimating dependences. Thus we shall bound the rate of uniform convergence of the means to mathematical expectations under the condition (7.15) and the condition that the class of functions possesses a bounded capacity in any one of the above-stated senses.

† The mean of the pth order of a random variable ξ is defined as $\sqrt[p]{M\xi^p}$.

§6 Two Theorems on Uniform Convergence

In this section we shall prove two theorems which bound the rate of uniform convergence of the means to the mathematical expectations. We shall consider the case when the set of functions $F(x, \alpha)$ consists of a finite number of elements and the case when the set of functions can be covered by a finite ε-net in either the C or the L_p^2 metric.

The proof of both theorems rely heavily on the following fact: let a function $F(x, \alpha^*)$ be such that the condition

$$\frac{\sqrt[p]{\int\int (y - F(x, \alpha^*))^{2p} P(x, y)\, dx\, dy}}{\int\int (y - F(x, \alpha^*))^2 P(x, y)\, dx\, dy} \leq \tau, \qquad p > 1 \tag{7.16}$$

is satisfied. Then if restriction (7.16) is stipulated for $p > 2$, the inequality

$$P\left\{\frac{I(\alpha^*) - I_{\text{emp}}(\alpha^*)}{I(\alpha^*)} > \tau a(p) \varkappa\right\} < 24 l e^{-\varkappa^2 l/4} \tag{7.17}$$

is valid, where

$$a(p) = \sqrt[p]{\frac{(p-1)^{p-1}}{2(p-2)^{p-1}}}. \tag{7.18}$$

If restriction (7.16) is stipulated for $1 < p \leq 2$, then the inequality

$$P\left\{\frac{I(\alpha^*) - I_{\text{emp}}(\alpha^*)}{I(\alpha^*)} > \tau V_p(\varkappa)\right\} < 24 l \exp\left\{-\frac{\varkappa^2}{4} l^{2-(2/p)}\right\} \tag{7.19}$$

where

$$V_p(\varkappa) = \varkappa \sqrt[p]{\left(1 - \frac{\ln \varkappa}{\sqrt[p-1]{p(p-1)}}\right)^{p-1}}$$

holds. Note that for $p > 3$ the values of $a(p)$ in (7.18) is close to 1. A large value for $a(p)$ is attained only when p is close to 2.

These bounds will be obtained as a corollary of Theorem 7.6 presented in Section 7.

Theorem 7.4. *Let the condition* (7.15) *be fulfilled, and the class of functions* $F(x, \alpha)$ *consist of a finite number* N *of elements. Then under* (7.15) *with* $p > 2$, *the inequality*

$$I(\alpha) \leq \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau a(p) \sqrt{\dfrac{\ln N + \ln l - \ln(\eta/24)}{l}}}\right]_{\infty} \tag{7.20}$$

is fulfilled with probability $1 - \eta$ *simultaneously for all functions in the class* $F(x, \alpha)$; *if, however* $1 < p \leq 2$, *then the inequality*

$$I(\alpha) \leq \left[\frac{I_{\text{emp}}(\alpha)}{1 - \tau V_p\left(2\sqrt{\dfrac{\ln N + \ln l - \ln(\eta/24)}{l^{2-(2/p)}}}\right)} \right]_\infty, \tag{7.21}$$

where

$$V_p(\varkappa) = (\varkappa)\sqrt[p]{\left(1 - \frac{\ln \varkappa}{\sqrt[p-1]{p(p-1)}}\right)^{p-1}},$$

$$[z]_\infty = \begin{cases} z & \text{for } z \geq 0, \\ \infty & \text{for } z < 0, \end{cases}$$

is fulfilled with probability $1 - \eta$ *simultaneously for all functions* $F(x, \alpha)$.

PROOF. Let $p > 2$ in the condition (7.15). We utilize the inequality

$$P\left\{\sup_i \frac{I(\alpha_i) - I_{\text{emp}}(\alpha_i)}{I(\alpha_i)} > \varkappa\tau a(p)\right\} < N \sup_i P\left\{\frac{I(\alpha_i) - I_{\text{emp}}(\alpha_i)}{I(\alpha_i)} > \varkappa\tau a(p)\right\}. \tag{7.22}$$

We bound the second factor on the right-hand side of (7.22) using (7.17). We thus obtain

$$P\left\{\sup_i \frac{I(\alpha_i) - I_{\text{emp}}(\alpha_i)}{I(\alpha_i)} > \tau\varkappa a(p)\right\} < 24Nle^{-\varkappa^2 l/4},$$

which can be written in the following equivalent form: with probability $1 - \eta$ the inequalities

$$I(\alpha_i) \leq \left[\frac{I_{\text{emp}}(\alpha_i)}{1 - 2\tau a(p)\sqrt{\dfrac{\ln N + \ln l + \ln(\eta/24)}{l}}} \right]_\infty$$

are valid simultaneously for all α_i. The first assertion of the theorem is proved.

Analogously in the case $1 < p \leq 2$ we shall use the bound (7.19). Applying this bound to the right-hand side of (7.22), we obtain a bound on the rate of uniform convergence which is equivalent to the assertion of the theorem. □

Theorem 7.5. *Let the condition* (7.15) *be satisfied, and let the set* $F(x, \alpha)$ *be covered by a finite* ε*-net. Then one can assert with probability* $1 - \eta$ *that the*

quality of the function $F(x, \alpha_{\text{emp}})$ *which yields the minimum for the empirical risk is bounded by*

$$I(\alpha_{\text{emp}}) \leq \left(\varepsilon + \sqrt{\varepsilon^2 + \left[\frac{I_{\text{emp}}(\alpha_i(\alpha_{\text{emp}}))}{1 - T(\varepsilon)}\right]_\infty}\right)^2,$$

where $F(x, \alpha_i(\alpha_{\text{emp}}))$ *is an element of the* ε*-net closest to* $F(x, \alpha_{\text{emp}})$,

$$T(\varepsilon)\begin{cases} = 2\tau a(p)\sqrt{\dfrac{\ln N(\varepsilon) + \ln l - \ln(\eta/24)}{l}} & \text{for } p > 2; \\ = \tau V_p\left(2\sqrt{\dfrac{\ln N(\varepsilon) + \ln l - \ln(\eta/24)}{l^{2-(2/p)}}}\right) & \text{for } 1 < p \leq 2. \end{cases}$$

Remark. Theorem 7.5 is valid for any ε, which defines a ε-net chosen *a priori*, i.e., before the sample is taken.

In particular ε may be chosen from the condition of the minimum for the expression

$$\varepsilon + \sqrt{\varepsilon^2 + \left[\frac{c}{1 - T(\varepsilon)}\right]_\infty},$$

where c is a constant. It is reasonable to choose c to be close to the minimum of functional $I(\alpha_0)$. Thus *a priori* information on the value of $I(\alpha_0)$ is utilized for choosing an appropriate ε.

The *proof* of this theorem is basically analogous to the proof of Theorem 7.2.

(1) We choose an arbitrary ε-net. For $p > 2$, in view of Theorem 7.4, the inequality

$$I(\alpha_i) \leq \left[\frac{I_{\text{emp}}(\alpha_i)}{1 - 2\tau a(p)\sqrt{\dfrac{\ln N(\varepsilon) + \ln l - \ln(\eta/24)}{l}}}\right]_\infty \tag{7.23}$$

is satisfied with probability $1 - \eta$ simultaneously for all elements of the ε-net.

(2) In view of the bound (7.11) obtained in the proof of Theorem 7.2, the values of the functionals $I(\alpha)$ for functions $F(x, \alpha_{\text{emp}})$ and $F(x, \alpha_i(\alpha_{\text{emp}}))$ which are separated in either the C or the L_p^2 metric by an amount smaller than ε, differ by an amount not exceeding

$$|I(\alpha_{\text{emp}}) - I(\alpha_i(\alpha_{\text{emp}}))| < 2\varepsilon\sqrt{\max(I(\alpha_{\text{emp}}), I(\alpha_i(\alpha_{\text{emp}})))}. \tag{7.24}$$

(3) We shall consider two cases: $I(\alpha_{\text{emp}}) > I(\alpha_i(\alpha_{\text{emp}}))$ and $I(\alpha_{\text{emp}}) \leq I(\alpha_i(\alpha_{\text{emp}}))$. In the first case it follows from (7.23) and (7.24) that the bound

$$I(\alpha_{\text{emp}}) \leq \left[\frac{I_{\text{emp}}(\alpha_i(\alpha_{\text{emp}}))}{1 - 2\tau a(p)\sqrt{\dfrac{\ln N(\varepsilon) + \ln l - \ln(\eta/24)}{l}}}\right]_\infty + 2\varepsilon\sqrt{I(\alpha_{\text{emp}})} \tag{7.25}$$

is valid with probability $1 - \eta$. In the second case we have the bound

$$I(\alpha_{\text{emp}}) \leq \left[\frac{I_{\text{emp}}(\alpha_i(\alpha_{\text{emp}}))}{1 - 2\tau a(p) \sqrt{\dfrac{\ln N(\varepsilon) + \ln l - \ln(\eta/24)}{l}}} \right]_{\infty} + 2\varepsilon \sqrt{I(\alpha_i(\alpha_{\text{emp}}))} \tag{7.25a}$$

with the same probability.

(4) Solving the inequality (7.25) for $I(\alpha_{\text{emp}})$ we obtain

$$I(\alpha_{\text{emp}}) \leq \left(\varepsilon + \sqrt{\varepsilon^2 + \left[\frac{I_{\text{emp}}(\alpha_i(\alpha_{\text{emp}}))}{1 - 2\tau a(p) \sqrt{\dfrac{\ln N(\varepsilon) + \ln l - \ln(\eta/24)}{l}}} \right]_{\infty}} \right)^2. \tag{7.26}$$

Taking (7.23) into account we verify that the bound (7.26) is valid also in the case (7.25a).

The theorem for the case $1 < p \leq 2$ is proved in the same manner. □

Remark. As in the case in Theorem 7.2, the bound (7.26) will be smaller ($N(\varepsilon)$ is smaller) provided the ε-net is constructed in the L_p^2 metric, i.e., when the information about the density $P(x)$ is utilized.

§7 Theorem on Uniform Relative Deviation

We now prove the basic theorem.

Theorem 7.6. *Let the condition* (7.15) *be satisfied and the set of functions* $F(x, \alpha)$ *possess a finite capacity* $h < l$; *then if* $p > 2$, *the inequality*

$$I(\alpha) \leq \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau a(p) \sqrt{\dfrac{h\left(\ln \dfrac{2l}{h} + 1\right) - \ln \dfrac{\eta}{12}}{l}}} \right]_{\infty},$$

where

$$a(p) = \sqrt[p]{\left(\frac{p-1}{p-2}\right)^{p-1} \cdot \frac{1}{2}}$$

is fulfilled with probability $1 - \eta$ *simultaneously for all functions* $F(x, \alpha)$; *if however* $1 < p \leq 2$, *the inequality*

$$I(\alpha) \leq \left[\frac{I_{\text{emp}}(\alpha)}{1 - \tau V_p\left(2\sqrt{\dfrac{h\left(\ln \dfrac{2l}{h} + 1\right) - \ln \dfrac{\eta}{12}}{l^{2-(2/p)}}}\right)} \right]_{\infty},$$

where

$$V_p(\varkappa) = \varkappa \sqrt[p]{\left(1 - \frac{\ln \varkappa}{\sqrt[p-1]{p(p-1)}}\right)^{p-1}}$$

is fulfilled with probability $1 - \eta$ *simultaneously for all functions* $F(x, \alpha)$.

We prove the theorem first for the case $p > 2$ and then for $1 < p \leq 2$.

To begin with we express the functional $I(\alpha)$ in terms of the Lebesgue integral

$$I(\alpha) = \int_0^\infty P\{(y - F(x, \alpha))^2 > t\}\, dt. \tag{7.27}$$

Observe that for any fixed α and arbitrary t the probability of the event $\{(y - F(x, \alpha))^2 > t\}$ is expressed in terms of the distribution function of a positive random variable $\xi(\alpha) = (y - F(x, \alpha))^2$; namely, the cumulative distribution function of $\xi(\alpha)$,

$$\Phi(\xi(\alpha) \leq t) = \Phi_\alpha(t),$$

is related to the probability of occurrence of event $\{(y - F(x, \alpha))^2 > t\}$ as follows:

$$P\{(y - F(x, \alpha))^2 > t\} = 1 - \Phi_\alpha(t).$$

Thus the functional (7.27) can be written in the form

$$I(\alpha) = \int (1 - \Phi_\alpha(t))\, dt.$$

We introduce a new functional

$$R(\alpha) = \int \sqrt{1 - \Phi_\alpha(t)}\, dt.$$

It is easy to see that this functional exceeds $I(\alpha)$, since

$$1 - \Phi_\alpha(t) < \sqrt{1 - \Phi_\alpha(t)}.$$

The following lemma is valid.

Lemma. *If for each function of the set $F(x, \alpha)$ the functional $R(\alpha)$ exists and the set of functions $F(x, \alpha)$ has a finite capacity $h < l$, then the inequality*

$$P\left\{\sup_{\alpha} \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{R(\alpha)} > \varkappa\right\} < 8m^S(2l)e^{-\varkappa^2 l/4} < 12\frac{(2l)^h}{h!}e^{-\varkappa^2 l/4} \tag{7.28}$$

is valid.

PROOF. Denote by $A_{\alpha,i}$ the event $\{(y - F(x, \alpha))^2 > i/n\}$. Consider the expression

$$\frac{I(\alpha) - I_{\text{emp}}(\alpha)}{R(\alpha)} = \frac{\lim_{n\to\infty}\left[\sum_{i=1}^{\infty}\frac{1}{n}P(A_{\alpha,i}) - \sum_{i=1}^{\infty}\frac{1}{n}\nu(A_{\alpha,i})\right]}{R(\alpha)}. \tag{7.29}$$

We show that if the inequality

$$\sup_{\alpha,i}\frac{P(A_{\alpha,i}) - \nu(A_{\alpha,i})}{\sqrt{P(A_{\alpha,i})}} \le \varkappa \tag{7.30}$$

is fulfilled, then the inequality

$$\sup_{\alpha}\frac{I(\alpha) - I_{\text{emp}}(\alpha)}{R(\alpha)} \le \varkappa$$

is fulfilled as well. Indeed, (7.29) and (7.30) imply that

$$\sup_{\alpha}\frac{I(\alpha) - I_{\text{emp}}(\alpha)}{R(\alpha)} \le \sup_{\alpha}\frac{\lim_{n\to\infty}\varkappa\sum_{i=1}^{\infty}\frac{1}{n}\sqrt{P(A_{\alpha,i})}}{R(\alpha)} = \sup_{\alpha}\frac{\varkappa R(\alpha)}{R(\alpha)} = \varkappa.$$

Thus the probability that the inequality

$$\sup_{\alpha}\frac{I(\alpha) - I_{\text{emp}}(\alpha)}{R(\alpha)} > \varkappa$$

is valid does not exceed the corresponding probability for the validity of

$$\sup_{\alpha,i}\frac{P(A_{\alpha,i}) - \nu(A_{\alpha,i})}{\sqrt{P(A_{\alpha,i})}} > \varkappa.$$

On the other hand, in view of Theorem A.3 of the Appendix to Chapter 6, the bound

$$P\left\{\sup_{\alpha,i}\frac{P(A_{\alpha,i}) - \nu(A_{\alpha,i})}{\sqrt{P(A_{\alpha,i})}} > \varkappa\right\} < 8m^S(2l)e^{-\varkappa^2 l/4}.$$

holds, which implies that

$$P\left\{\sup_{\alpha}\frac{I(\alpha)-I_{\text{emp}}(\alpha)}{R(\alpha)}>\varkappa\right\}<8m^{S}(2l)e^{-\varkappa^{2}l/4}. \tag{7.31}$$

Noting that $m^{S}(l)<1.5l^{n}/h!$, we arrive at the bound (7.28). The lemma is thus proved.

PROOF OF THE THEOREM. The statement of the lemma involves the following condition: for any function $F(x,\alpha)$ there exists a functional $R(\alpha)$. We now show that the functional $R(\alpha)$ exists provided the random variable $\xi(\alpha)=(y-F(x,\alpha))^{2}$ possesses a moment of order greater than second (even a noninteger one). Moreover for $p>2$ the relation

$$R(\alpha)<\sqrt[p]{M\xi^{p}(\alpha)}\cdot a(p),$$

where

$$a(p)=\sqrt[p]{\frac{(p-1)^{p-1}}{2(p-2)^{p-1}}},$$

is valid. Indeed, the transformation

$$M\xi^{p}(\alpha)=\int(y-F(x,\alpha))^{2p}P(x,y)\,dx\,dy$$

$$=\int_{0}^{\infty}t^{p}\,d\Phi_{\alpha}(t)=p\int_{0}^{\infty}t^{p-1}(1-\Phi_{\alpha}(t))\,dt$$

is valid. On the other hand, by definition

$$R(\alpha)=\int_{0}^{\infty}\sqrt{1-\Phi_{\alpha}(t)}\,dt.$$

Now let the pth moment be $m_{p}(\alpha)$:

$$p\int_{0}^{\infty}t^{p-1}(1-\Phi_{\alpha}(t))\,dt=m_{p}(\alpha).$$

We shall obtain a distribution $\Phi_{\alpha}(t)$ such that $R(\alpha)$ is maximized.

For this purpose we construct the Lagrange function

$$L(\alpha)=R(\alpha)-\lambda M\xi^{p}(\alpha)$$

$$=\int_{0}^{\infty}\sqrt{1-\Phi_{\alpha}(t)}\,dt-\lambda p\int_{0}^{\infty}t^{p-1}(1-\Phi_{\alpha}(t))\,dt. \tag{7.32}$$

We determine a probability distribution function $\Phi_{\alpha}(t)$ for which the maximum of $L(\alpha)$ is obtained. Denote $z^{2}=1-\Phi_{\alpha}(t)$, $b=\lambda p$, and rewrite (7.32) using this notation:

$$L(\alpha)=\int_{0}^{\infty}z(1-bzt^{p-1})\,dt. \tag{7.33}$$

The function z at which the maximum of the functional (7.33) is attained is defined by

$$1 - 2bzt^{p-1} = 0,$$

which implies that

$$z = \left(\frac{t_0}{t}\right)^{p-1},$$

where $t_0 = (1/2b)^{1/(1-p)}$.

Since $z(t)$ varies between 1 and 0 as t varies between 0 and ∞, the optimal function $z(t)$ is

$$z(t) = \begin{cases} 1 & \text{if } t < t_0, \\ \left(\dfrac{t_0}{t}\right)^{p-1} & \text{if } t \geq t_0. \end{cases}$$

We now compute $\max_\alpha R(\alpha)$ (recalling that $p > 2$):

$$\max_\alpha R(\alpha) = \int_0^\infty z(t)\,dt = t_0 + \int_0^\infty \left(\frac{t_0}{t}\right)^{p-1} dt = \frac{p-1}{p-2}\,t_0. \tag{7.34}$$

On the other hand, express t_0 in terms of m_p:

$$\begin{aligned} m_p(\alpha) &= p\int_0^\infty z^2(t)t^{p-1}\,dt \\ &= p\int_0^{t_0} t^{p-1}\,dt + p\int_{t_0}^\infty \left(\frac{t_0}{t}\right)^{2p-2} t^{p-1}\,dt = 2t_0^p\left(\frac{p-1}{p-2}\right). \end{aligned} \tag{7.35}$$

Substituting the value of t_0 obtained from (7.35) into (7.34), we arrive at

$$\sup_\alpha \frac{R(\alpha)}{\sqrt[p]{m_p(\alpha)}} = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}} = a(p),$$

which implies that for $p > 2$

$$R(\alpha) < \sqrt[p]{M\xi^p(\alpha)}\,a(p). \tag{7.36}$$

Utilizing the lemma and the bound (7.36), we prove the first part of the theorem. Note that under the conditions of the theorem the inequality

$$R(\alpha) < \tau a(p) I(\alpha) \tag{7.37}$$

is valid. We utilize the bound (7.37) to improve the inequality (7.28):

$$\begin{aligned} &P\left\{\sup_\alpha \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{I(\alpha)} > \tau a(p)\varkappa\right\} \\ &\qquad < P\left\{\sup_\alpha \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{R(\alpha)} > \varkappa\right\} < 12\,\frac{(2l)^h}{h!}\,e^{-\varkappa^2 l/4}. \end{aligned} \tag{7.38}$$

The first assertion of the theorem is equivalent to this inequality.

We now prove the second part of the theorem. Consider the difference

$$I(\alpha) - I_{\text{emp}}(\alpha) = \lim_{n \to \infty} \sum_{i=1}^{\infty} \frac{1}{n} (P(A_{\alpha,i}) - \nu(A_{\alpha,i})). \tag{7.39}$$

Assume that for all events $A_{\alpha,i}$ the condition

$$P(A_{\alpha,i}) - \nu(A_{\alpha,i}) \leq \varkappa \sqrt[p]{P(A_{\alpha,i})} \tag{7.40}$$

is fulfilled. Moreover the inequality

$$P(A_{\alpha,i}) - \nu(A_{\alpha,i}) \leq P(A_{\alpha,i}) \tag{7.41}$$

is always valid. To compute the sum (7.39) we apply the bound (7.40) to the summands corresponding to the events $A_{\alpha,i}$ for which $P(A_{\alpha,i}) > \varkappa^{p/(p-1)}$. For the summands for which the events $A_{\alpha,i}$ satisfy $P(A_{\alpha,i}) \leq \varkappa^{p/(p-1)}$ we shall utilize the trivial bound (7.41). We thus obtain

$$I(\alpha) - I_{\text{emp}}(\alpha)$$

$$\leq \varkappa \int_{1-\Phi_\alpha(t) > \varkappa^{p/(p-1)}} \sqrt[p]{1 - \Phi_\alpha(t)}\, dt + \int_{1-\Phi_\alpha(t) \leq \varkappa^{p/(p-1)}} (1 - \Phi_\alpha(t))\, dt. \tag{7.42}$$

We now find the maximal value (with respect to $\Phi_\alpha(t)$) of the right-hand side of the inequality under the condition that the pth moment takes on some fixed value m_p, i.e.,

$$p \int_0^\infty t^{p-1}(1 - \Phi_\alpha(t))^p\, dt = m_p$$

For this purpose we again use the method of Lagrange multipliers, denoting

$$z^p = 1 - \Phi_\alpha(t).$$

We thus seek the maximum of the expression

$$L(\alpha) = \int_{z > \varkappa^{-p+1}} \varkappa z\, dt + \int_{z \leq \varkappa^{-p+1}} z^p\, dt - \lambda \int_0^\infty t^{p-1} z^p\, dt.$$

Represent $L(\alpha)$ in the form

$$L(\alpha) = \int_{z > \varkappa^{-p+1}} (\varkappa z - \lambda t^{p-1} z^p)\, dt + \int_{z \leq \varkappa^{-p+1}} (z^p - \lambda t^{p-1} z^p)\, dt,$$

where the first summand defines the function $z(t)$ in the domain $z > \varkappa$ and the second in the domain $z \leq \varkappa$. The first summand attains its absolute maximum at

$$z = \sqrt[p-1]{\frac{\varkappa}{p\lambda} \frac{1}{t}}.$$

However, taking into account that z is a monotonically decreasing function from 1 to $\varkappa$, we obtain

$$z(t) = \begin{cases} 1 & \text{if} \quad 0 \le t < \sqrt[p-1]{\dfrac{\varkappa}{p\lambda}}, \\ \sqrt[p-1]{\dfrac{\varkappa}{p\lambda}}\,\dfrac{1}{t} & \text{if} \quad \sqrt[p-1]{\dfrac{\varkappa}{p\lambda}} \le t < \sqrt[p-1]{\dfrac{1}{p\lambda}}. \end{cases}$$

The second summand attains its maximum in the doamin $z \le \varkappa^{p+1}$ for the function

$$z(t) = \begin{cases} \sqrt[p-1]{\varkappa} & \text{if} \quad \sqrt[p-1]{\dfrac{1}{p\lambda}} \le t < \sqrt[p-1]{\dfrac{1}{\lambda}}, \\ 0 & \text{if} \quad t \ge \sqrt[p-1]{\dfrac{1}{\lambda}}. \end{cases}$$

We thus finally obtain

$$z(t) = \begin{cases} 1 & \text{if} \quad 0 \le t < \sqrt[p-1]{\dfrac{\varkappa}{p\lambda}}, \\ \sqrt[p-1]{\dfrac{\varkappa}{p.\lambda}}\,\dfrac{1}{t} & \text{if} \quad \sqrt[p-1]{\dfrac{\varkappa}{p\lambda}} \le t < \sqrt[p-1]{\dfrac{1}{p\lambda}}, \\ \sqrt[p-1]{\varkappa} & \text{if} \quad \sqrt[p-1]{\dfrac{1}{p\lambda}} \le t < \sqrt[p-1]{\dfrac{1}{\lambda}}, \\ 0 & \text{if} \quad \sqrt[p-1]{\dfrac{1}{\lambda}} \le t < \infty. \end{cases}$$

We now express the pth moment m_p in terms of the Lagrange multiplier λ. For this purpose we compute the pth moment

$$m_p = p\int_0^\infty t^{p-1}z^p\,dt = \left(\frac{\varkappa}{\lambda}\right)^{p/(p-1)}\left(1 - \frac{\ln \varkappa}{\sqrt[p-1]{p}(p-1)}\right). \tag{7.43}$$

Analogously we compute the quantity

$$I(\alpha) - I_{\text{emp}}(\alpha) \le \varkappa\int_0^{\sqrt[p-1]{1/p\lambda}} z\,dt + \int_{\sqrt[p-1]{1/p\lambda}}^\infty z^p\,dt = \varkappa\left(\frac{\varkappa}{\lambda}\right)^{1/(p-1)}\left(1 - \frac{\ln x}{\sqrt[p-1]{p}(p-1)}\right). \tag{7.43a}$$

It follows from (7.43) and (7.43a) that

$$\sup_\alpha \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{\sqrt[p]{m_p(\alpha)}} < V_p(\varkappa), \tag{7.44}$$

where

$$V_p(\varkappa) = \varkappa\sqrt[p]{\left(1 - \frac{\ln \varkappa}{\sqrt[p-1]{p}(p-1)}\right)^{p-1}}.$$

Thus we have shown that the condition (7.40) implies the inequality (7.44). Therefore the probability of the event

$$\left\{\sup_{\alpha} \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{\sqrt[p]{m_p(\alpha)}} > V_p(\varkappa)\right\}$$

does not exceed the probability of the event

$$\left\{\sup_{\alpha, i} \frac{P(A_{\alpha, i}) - \nu(A_{\alpha, i})}{\sqrt[p]{P(A_{\alpha, i})}} > \varkappa\right\}.$$

According to the assertion of Theorem A.3 in the Appendix to Chapter 6, the probability of this event for $l > h$ is bounded by (A.16); this implies that

$$P\left\{\sup_{\alpha} \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{\sqrt[p]{m_p(\alpha)}} > V_p(\varkappa)\right\} < 12 \frac{(2l)^h}{h!} \exp\left\{-\frac{\varkappa^2}{4} l^{2-(2/p)}\right\}.$$

On the other hand, in view of the condition of the theorem (Equation (7.15)),

$$\sqrt[p]{m_p(\alpha)} \leq \tau I(\alpha).$$

Taking this into account, we obtain

$$P\left\{\sup_{\alpha} \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{I(\alpha)} > \tau V_p(\varkappa)\right\} < P\left\{\sup_{\alpha} \frac{I(a) - I_{\text{emp}}(\alpha)}{\sqrt[p]{m_p(\alpha)}} > V_p(\varkappa)\right\}.$$

We thus finally arrive at the inequality

$$P\left\{\sup_{\alpha} \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{I(\alpha)} > \tau V_p(\varkappa)\right\} < 12 \frac{(2l)^h}{h!} \exp\left\{-\frac{\varkappa^2}{4} l^{2-(2/p)}\right\} \quad (7.45)$$

for $l > h$. This inequality is equivalent to the assertion of the second part of the theorem. □

Remark. For the proofs of Theorems 7.4 and 7.5 we have utilized bounds on relative deviations, (7.17) and (7.19). These bounds may be easily obtained from the inequalities (7.38) and (7.45), taking into account that the capacity of the class of decision rules $F(x, \alpha)$ formed by a fixed function $F(x, \alpha^*)$ equals 1.

§8 Remarks on a General Theory of Risk Estimation

We have thus constructed a theory of uniform convergence of the means to their mathematical expectations. Formally this theory was constructed for quadratic loss functions. However, the results obtained are also valid for general loss functions.

Below we state the basic assertions of the theory of uniform deviations of empirical estimators from the means in a general setup. The proofs of these assertions are identical to the proofs of the analogous theorems considered above.

Let $Q(z, \alpha)$ be a parametric family of nonnegative functions satisfying the following conditions:

(1) for any fixed value of the parameter $\alpha^* \in \Lambda$ the functions $Q(z, \alpha)$ are measurable in z;
(2) the set of functions $Q(z, \alpha)$ has a finite capacity h (the indicator functions $\theta(Q(z, \alpha) + \beta)$ have a finite capacity h).

Then the following assertions on the rate of uniform convergence of empirical means

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha),$$

constructed from a sample $z_1, \ldots, z_l$ to their mathematical expectations

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz$$

are valid.

Assertion 1. *If for functions $Q(z, \alpha)$ the functional*

$$R_p(\alpha) = \int \sqrt[p]{1 - P\{Q(z, \alpha) \leq t\}}\, dt$$

exists, then for $l > h$ the inequality

$$P\left\{\sup_{\alpha} \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{R_p(\alpha)} > \varkappa\right\} \begin{cases} < 12 \dfrac{(2l)^h}{h!} \exp\left\{-\dfrac{\varkappa^2}{4} l^{2-(2/p)}\right\} & \text{for } 1 < p \leq 2, \\ < 12 \dfrac{(2l)^h}{h!} \exp\left\{-\dfrac{\varkappa^2}{b(p)} l\right\} & \text{for } p > 2, \end{cases} \tag{7.46}$$

where

$$b(p) = \sqrt[p]{4\left(\frac{p}{p-1}\right)^p \left(\frac{p-2}{p-1}\right)^{p-2}}$$

is valid.

Assertion 2. *If for functions* $Q(z, \alpha)$ *the pth moment* $(1 < p \leq 2)$

$$m_p(\alpha) = \int Q^p(z, \alpha) P(z)\, dz$$

exists, then the inequality

$$P\left\{\sup_\alpha \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{\sqrt{m_p(\alpha)}} > \varkappa \sqrt[p]{\left(1 - \frac{\ln \varkappa}{\sqrt[p-1]{p(p-1)}}\right)^{p-1}}\right\} < 12 \frac{(2l)^h}{h!} \exp\left\{-\frac{\varkappa^2}{4} l^{2-(2/p)}\right\}$$

is valid for $l > h$.

Assertion 3. *If for functions* $Q(z, \alpha)$ *the pth moment* $(p > 2)$

$$m_p(\alpha) = \int Q^p(z, \alpha) P(z)\, dz$$

exists, then for $l > h$ *we have the inequality*

$$P\left\{\sup_\alpha \frac{I(\alpha) - I_{\text{emp}}(\alpha)}{\sqrt[p]{m_p(\alpha)}} > a(p)\varkappa\right\} < 12 \frac{(2l)^h}{h!} e^{-\varkappa^2 l/4},$$

where

$$a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}.$$

Assertion 4. *If the condition*

$$\sup_\alpha \frac{\sqrt[p]{m_p(\alpha)}}{I(\alpha)} \leq \tau$$

is fulfilled for $p > 2$, *then for* $l > h$ *the inequality*

$$I(\alpha) \leq \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau a(p)\sqrt{\dfrac{h\left(\ln \dfrac{2l}{h} + 1\right) - \ln \dfrac{\eta}{12}}{l}}}\right]_\infty \tag{7.47}$$

is satisfied with probability $1 - \eta$ *simultaneously for all* α. *If, however, the condition*

$$\sup_\alpha \frac{\sqrt[p]{m_p(\alpha)}}{I(\alpha)} \leq \tau$$

is fulfilled for $1 < p \leq 2$, *then for all* $l > h$ *the inequality*

$$I(\alpha) \leq \left[\frac{I_{\text{emp}}(\alpha)}{1 - \tau V_p \left(2 \sqrt{\frac{h\left(\ln \frac{2l}{h} + 1 \right) - \ln \frac{\eta}{12}}{l^{2-(2/p)}}} \right)} \right]_{\infty} \tag{7.48}$$

is satisfied with probability $1 - \eta$ *simultaneously for all* α, *where*

$$V_p(\varkappa) = \varkappa \sqrt[p]{\left(1 - \frac{\ln \varkappa}{\sqrt[p-1]{p(p-1)}} \right)^{p-1}}.$$

In Chapters 8 and 9 we shall utilize the theory of uniform convergence developed herein to construct extremal algorithms for estimating dependences in the case of samples of finite sizes. Here we shall note that if the condition (7.15) is satisfied and the capacity of the class of functions $F(x, \alpha)$ is bounded, then according to the theory described the method of minimizing empirical risk leads us to the determination of a function which is close to the best in the class (provided the sample size is sufficiently large). Indeed, in this case the denominator in the bounds (7.47) and (7.48) is close to 1 and the value of the expected risk determines the value of the empirical risk.

Appendix to Chapter 7

Theory of Uniform Convergence of Means to Their Mathematical Expectations: Necessary and Sufficient Conditions

§A1 ε-entropy

In the Appendix to Chapter 6 sufficient conditions for the uniform convergence of frequencies to probabilities were established. These conditions are sufficient in order that the equality

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left|MF(x,\alpha) - \frac{1}{l}\sum_{i=1}^{l} F(x_i,\alpha)\right| > \varepsilon\right\} = 0 \qquad \text{(A.1)}$$

be fulfilled for a given set of indicator functions $F(x, \alpha)$, $\alpha \in \Lambda$ as the sample size of a random indepedent sample of vectors $x_1, \ldots, x_l$ increases.

In this Appendix we shall indicate necessary and sufficient conditions for the uniform convergence of means to their mathematical expectations in the case of uniformly bounded families of functions

$$0 \le F(x,\alpha) \le C, \qquad \alpha\in\Lambda. \qquad \text{(A.2)}$$

(These are conditions which are necessary and sufficient for the fulfillment of the equality (A.1) for the family (A.2).) Below we shall assume without loss of generality that $C = 1$.† To state these conditions precisely we introduce several notions.

Let A be a bounded set of vectors in E_l. A finite set $T \subset E_l$ such that for any $y \in A$ there exists an element $t \in T$ satisfying $\rho(t, y) < \varepsilon$ is called a *relative ε-net* of A in E_l.

Below we shall assume that the metric is defined by

$$\rho(t,y) = \max_{1\le i\le n} |t^i - y^i|, \qquad t = (t^1,\ldots,t^n), \quad y = (y^1,\ldots,y^n),$$

and the norm of a vector z is given by $\|z\| = \max_{1\le i\le n} |z^i|$.

† Note that indicator functions satisfy the condition (A.2).

If an ε-net T of a set A is such that $T \subset A$, then we call it a *proper ε-net* of the set A.

The minimal number of elements in an ε-net of the set A relative to E_l will be denoted by $N(\varepsilon, A)$, the minimal number of elements in a proper ε-net is denoted by $N_0(\varepsilon, A)$. It is easy to see that

$$N_0(\varepsilon, A) \geq N(\varepsilon, A). \tag{A.3}$$

On the other hand

$$N_0(2\varepsilon, A) < N(\varepsilon, A). \tag{A.4}$$

Indeed, let T be a minimal ε-net of A relative to E_l. We assign to each element $t \in T$ an element $y \in A$ such that $\rho(t, y) < \varepsilon$ (such an element y always exists, since otherwise the ε-net could have been reduced). The totality T_0 of elements of this kind forms a proper 2ε-net in A (for each $y \in A$ there exists $t \in T$ such that $\rho(y, t) < \varepsilon$, and for such a $t \in T$ there exists $\tau \in T_0$ such that $\rho(t, \tau) < \varepsilon$ and hence $\rho(y, \tau) < 2\varepsilon$).

Let $F(x, \alpha)$ be a class of numerical functions in the variable $x \in X$ depending on parameter $\alpha \in \Lambda$. Let $x_1, \ldots, x_l$ be a sample. Consider in the space E_l a set A of vectors z with coordinates $z^i \in F(x_i, \alpha)$, $i = 1, \ldots, l$, formed by all $\alpha \in \Lambda$.

If the condition $0 \leq F(x, \alpha) \leq 1$ is fulfilled, then the set $A = A(x_1, \ldots, x_l)$ belongs to an l-dimensional cube $0 \leq z^i \leq 1$ and is therefore bounded and possesses a finite ε-net. The number of elements of a minimal relative ε-net of A in E_l is $N(\varepsilon; A(x_1, \ldots, x_l)) = N^{\Lambda}(x_1, \ldots, x_l; \varepsilon)$. The number of elements of a minimal proper ε-net is $N_0^{\Lambda}(x_1, \ldots, x_l; \varepsilon)$. If a probability measure P_X is defined on X and $x_1, \ldots, x_l$ is an independent random sample and $N^{\Lambda}(x_1, \ldots, x_l; \varepsilon)$ is a function measurable with respect to this measure on sequences $x_1, \ldots, x_l$ then there exists an average ε-entropy (or simply an ε-entropy)

$$H^{\Lambda}(\varepsilon, l) = M \log_2 N^{\Lambda}(x_1, \ldots, x_l; \varepsilon).$$

It is easy to verify that a minimal relative ε-net satisfies

$$N^{\Lambda}(x_1, \ldots, x_{l+k}; \varepsilon) \leq N^{\Lambda}(x_1, \ldots, x_l; \varepsilon) N^{\Lambda}(x_{l+1}, \ldots, x_{l+k}; \varepsilon): \tag{A.5}$$

(Recall that

$$\rho(z_1, z_2) = \max_{1 \leq i \leq n} |z_1^i - z_2^i|).$$

Indeed, in this case a direct product of relative ε-nets is also a relative ε-net. Thus

$$H^{\Lambda}(\varepsilon, l + k) \leq H^{\Lambda}(\varepsilon, l) + H^{\Lambda}(\varepsilon, k). \tag{A.6}$$

In the end of this section it will be shown that there exists the limit

$$c(\varepsilon) = \lim_{l \to \infty} \frac{H^{\Lambda}(\varepsilon, l)}{l}, \qquad 0 \leq c(\varepsilon) \leq \log_2 \left[1 + \frac{1}{\varepsilon}\right]$$

and the convergence

$$\frac{\log_2 N^{\Lambda}(x_1, \ldots, x_l; \varepsilon)}{l} \xrightarrow[l\to\infty]{p} c(\varepsilon) \tag{A.7}$$

holds.

Consider two cases:

(1) $\lim_{l\to\infty} H^{\Lambda}(\varepsilon, l)/l = c(\varepsilon) = 0$ for all $\varepsilon > 0$.
(2) There exists an ε_0 such that $c(\varepsilon_0) > 0$ (then also for all $\varepsilon < \varepsilon_0$ the quantity $c(\varepsilon) > 0$).

It follows from (A.4) and (A.7) that in the first case

$$\lim_{l\to\infty} \frac{\log_2 N_0^{\Lambda}(x_1, \ldots, x_l; \varepsilon)}{l} \xrightarrow[l\to\infty]{p} 0 \tag{A.8}$$

for all $\varepsilon > 0$. It follows from (A.3) and (A.7) that in the second case

$$\lim_{l\to\infty} P\left\{\frac{\log_2 N_0^{\Lambda}(x_1, \ldots, x_l; \varepsilon)}{l} > c(\varepsilon_0) - \delta\right\} = 1 \tag{A.9}$$

for all $\varepsilon \le \varepsilon_0$, $\delta > 0$.

Below it will be shown that (A.8) implies uniform convergence of the means to their mathematical expectations, while under (A.9) such a convergence is not valid. Thus the following theorem is valid.

Theorem A.1. *The equality*

$$\lim_{l\to\infty} \frac{H^{\Lambda}(\varepsilon, l)}{l} = 0, \qquad \forall \varepsilon > 0$$

is a necessary and sufficient condition for the uniform convergence of means to their mathematical expectations for a bounded family of functions $F(x, \alpha)$, $\alpha \in \Lambda$.†

The next sections are devoted to the proof of this theorem.

We now prove (as in the information theory [65a]) that the limit (A.7) exists and the convergence (A.8) is valid.

1.1 Proof of the Existence of the Limit

As $0 \le H^{\Lambda}(\varepsilon, l)/l \le 1$, for any $\varepsilon_0 > 0$ there is a lower bound

$$\lim_{l\to\infty} \frac{H^{\Lambda}(\varepsilon_0, l)}{l} = c_0.$$

† For indicator functions $F(x, \alpha)$ we have $H^{\Lambda}(\varepsilon, l) \equiv M \log_2 \Delta^{\varepsilon}(x_1, \ldots, x_l)$ for all $0 < \varepsilon < 1$ (cf. Section A.2 of the Appendix to Chapter 6).

Therefore for any $\delta > 0$ such an l_0 can be found that

$$\frac{H^{\wedge}(\varepsilon_0, l_0)}{l_0} \leq c_0 + \delta.$$

Now take arbitrary $l > l_0$. Let $l = nl_0 + m$ where $n = [l/l_0]$. Then by virtue of (A.6)

$$\frac{H^{\wedge}(\varepsilon_0, l)}{l} = \frac{H^{\wedge}(\varepsilon_0, nl_0 + m)}{nl_0 + m} < \frac{nH^{\wedge}(\varepsilon_0, l_0) + m}{nl_0} < \frac{H^{\wedge}(\varepsilon_0, l_0)}{l_0} + \frac{1}{n}.$$

Strengthen the latter inequality

$$\frac{H^{\wedge}(\varepsilon_0, l)}{l} < \frac{H^{\wedge}(\varepsilon_0, l_0)}{l_0} + \frac{1}{n} < c_0 + \delta + \frac{1}{n}.$$

Since $n \to \infty$ when $l \to \infty$ we have

$$\overline{\lim_{l \to \infty}} \frac{H^{\wedge}(\varepsilon_0, l)}{l} \leq c_0 + \delta.$$

As $\delta > 0$ is arbitrary, the upper bound coincides with the lower one.

1.2 Proof of the Convergence of the Sequence

We prove that when l increases the sequence of random values

$$r^l = \frac{\log_2 N^{\wedge}(x_1, \ldots, x_l; \varepsilon_0)}{l}$$

converges in probability to the limit c_0. For this it is sufficient to show that for any $\delta > 0$

$$P_{\delta}^{+}(r^l) = P\{r^l > c_0 + \delta\} \xrightarrow[l \to \infty]{} 0$$

and for any $\mu > 0$

$$P_{\mu}^{-}(r^l) = P\{r^l < c_0 - \mu\} \xrightarrow[l \to \infty]{} 0.$$

Consider a random sequence

$$g_n^{l_0} = \frac{1}{n} \sum_{i=1}^{n} r_i^{l_0}$$

of independent random values r_i^l. Evidently

$$Mr^{l_0} = Mg_n^{l_0} = \frac{H^{\wedge}(\varepsilon_0, l_0)}{l_0}.$$

As $0 < r_l^{l_0} \leq 1$, we have

$$M(r^{l_0} - Mr^{l_0})^2 = D_2 \leq 1,$$

$$M(r^{l_0} - Mr^{l_0})^4 = D_4 \leq 1.$$

Therefore

$$M(g_n^{l_0} - Mg_n^{l_0})^4 = \frac{D_4}{n^3} + 3\frac{n+1}{n^3} D_2 < \frac{4}{n^2}.$$

Write the Chebyshev's inequality for the fourth moment

$$P\left\{\left|g_n^{l_0} - \frac{H^{\Lambda}(\varepsilon_0, l_0)}{l_0}\right| > \varkappa\right\} < \frac{4}{n^2\varkappa^4}.$$

Consider a random value g_n^l, where $l = nl_0 + m$. By virtue of (A.5)

$$r^l = r^{nl_0+m} \leq g_n^{l_0} + \frac{1}{n}.$$

Now let $\varkappa = \delta/3$, l_0 and $l = nl_0 + m$ be so large that

$$\frac{H^{\Lambda}(\varepsilon_0, l_0)}{l_0} - c_0 \leq \frac{\delta}{3},$$

$$\frac{1}{n} \leq \frac{\delta}{3}.$$

Then

$$P_{\delta}^{+}(r^l) = P\{r^l - c_0 > \delta\} \leq P\left\{\left|g_n^{l_0} - c_0 - \frac{2}{3}\delta\right| > \frac{\delta}{3}\right\} < \frac{244}{\delta^4 n^2}.$$

As $n \to \infty$ when $l \to \infty$

$$P_{\delta}^{+}(r^l) \underset{l\to\infty}{\longrightarrow} 0.$$

To bound the value $P_{\mu}^{-}(r^l)$ consider the equality

$$\int_0^{H^{\Lambda}(\varepsilon_0, l)/l} \left(\frac{H^{\Lambda}(\varepsilon_0, l)}{l} - r^l\right) dP(r^l) = \int_{H^{\Lambda}(\varepsilon_0, l)/l}^{1} \left(r^l - \frac{H^{\Lambda}(\varepsilon_0, l)}{l}\right) dP(r^l).$$

Mark its left part with R_1, the right one with R_2 and bound R_1 and R_2 for such l that

$$\frac{H^{\Lambda}(\varepsilon_0, l)}{l} - c_0 < \frac{\mu}{2}.$$

The lower bound of R_1 is

$$R_1 = \int_0^{H^{\Lambda}(\varepsilon_0, l)/l} \left(\frac{H^{\Lambda}(\varepsilon_0, l)}{l} - r^l\right) dP(r^l) \geq \frac{\mu}{2} \int_0^{c_0 - \mu} dP(r^l) = \frac{\mu}{2} P_{\mu}^{-}(r^l)$$

and the upper bound of R_2 is

$$R_2 = \int_{H^{\Lambda}(\varepsilon_0, l)/l}^{c_0+\delta} \left(r^l - \frac{H^{\Lambda}(\varepsilon_0, l)}{l}\right) dP(r^l) + \int_{c_0+\delta}^{1} \left(r^l - \frac{H^{\Lambda}(\varepsilon_0, l)}{l}\right) dP(r^l)$$

$$\leq \left|c_0 + \delta - \frac{H^{\Lambda}(\varepsilon_0, l)}{l}\right| + P_{\delta}^{+}(r^l).$$

Combining these bounds we obtain

$$\frac{\mu}{2} P_\mu^-(r^l) \leq \left| c_0 + \delta - \frac{H^\wedge(\varepsilon_0, l)}{l} \right| + P_\delta^+(r^l).$$

Since

$$\frac{H^\wedge(\varepsilon_0, l)}{l} \underset{l \to \infty}{\longrightarrow} c_0,$$

$$P_\delta^+(r^l) \underset{l \to \infty}{\longrightarrow} 0,$$

we obtain

$$\lim_{l \to \infty} P_\mu^-(r^l) \leq \frac{2\delta}{\mu}.$$

As δ and μ are arbitrary, we conclude that

$$P_\mu^-(r^l) \underset{l \to \infty}{\longrightarrow} 0.$$

§A2 The Quasicube

We shall define by induction an n-dimensional *quasicube* with an edge a.

Definition. A set Q in the space E_1 is called a one-dimensional quasicube with an edge a if Q is a segment $[c, c + a]$.

A set Q in the space E_n is called an n-dimensional quasicube with an edge a if there exists a coordinate subspace E_{n-1} (for simplicity it will be assumed below that this subspace is formed by the first $n - 1$ coordinates) such that a projection Q' of the set Q on this subspace is an $(n - 1)$-dimensional quasicube with an edge a and for each point $z_* \in Q^1$ $(z_* = (z_*^1, \ldots, z_*^{n-1}))$ the set of numerical values z^n such that $(z_*^1, \ldots, z_*^{n-1}, z^n) \in Q$ forms a segment $[c, c + a]$, where c in general does not depend on z_*.

The space E_{n-1} is called an $(n - 1)$-dimensional *canonical* space. In turn an $(n - 2)$-dimensional canonical space E_{n-2} can be constructed for this space and so on.

The totality of subspaces $E_1, \ldots, E_n$ is called a *canonical structure.*

The following lemma is valid.

Lemma A.1. *Let a convex set A belong to an l-dimensional cube whose coordinates satisfy*

$$0 \leq z^i \leq 1, \qquad i = 1, \ldots, l.$$

Let $V(A)$ be the l-dimensional volume of the set A.

If for some $1 \leq n \leq l, 0 \leq a \leq 1, l > 1$ *the condition*

$$V(A) > C_l^n a^{l-n} \tag{A.10}$$

is fulfilled, one can then find a coordinate n-dimensional subspace such that the projection of the set A on this subspace contains a quasicube with an edge a.

PROOF. We shall prove the lemma using an induction method.

(1) For $n = l$ the condition (A.10) is

$$V(A) > C_l^n = 1. \tag{A.11}$$

On the other hand

$$V(A) \leq 1. \tag{A.12}$$

Therefore the condition (A.1) is never fulfilled and the assertion of the lemma is trivially valid.

(2) For $n = 1$ and any l we shall prove the lemma by contradiction. Let there exist no one-dimensional coordinate space such that the projection of the set A on this space contains the segment $[c, c + a]$. The projection of a bounded convex set on the one-dimensional axis is either an open interval or a segment or a semiclosed interval. Consequently by assumption the length of this interval does not exceed a. However, then the set A itself is contained in an (ordinary) cube with an edge a. This implies that

$$V(A) \leq a^l.$$

Taking into account that $a \leq 1$, we obtain

$$V(A) < a^l < la^{l-1},$$

which contradicts the condition (A.10) of the lemma.

(3) Consider now the general inductive step. Let the lemma be valid for all $n < n_0$ for all l, and for $n = n_0 + 1$ for all l such that $n \leq l \leq l_0$. We shall show that it is valid for $n = n_0 + 1, l = l_0 + 1$.

Consider a coordinate subspace E_{l_0} of dimension l_0 consisting of vectors

$$z = (z^1, \ldots, z^{l_0}).$$

Let A^{I} be a projection of A on this subspace. (Clearly A^{I} is convex.)

If

$$V(A^{\mathrm{I}}) > C_{l_0}^n a^{l_0-n}, \tag{A.13}$$

then by the induction assumption there exists a subspace of dimension n such that the projection of the set A^{I} on this subspace contains a quasicube with an edge a. The lemma is thus proved in the case (A.13).

Let

$$V(A^{\mathrm{I}}) \leq C_{l_0}^n a^{l_0-n}. \tag{A.14}$$

Consider two functions

$$\varphi_1(z^1, \ldots, z^{l_0}) = \sup_z \{z: (z^1, \ldots, z^{l_0}, z) \in A\},$$

$$\varphi_2(z^1_1, \ldots, z^{l_0}) = \inf_z \{z: (z^1, \ldots, z^{l_0}, z) \in A\}.$$

These functions are convex upward and downward respectively. Therefore the function

$$\varphi_3(z^1, \ldots, z^{l_0}) = \varphi_1(z^1, \ldots, z^{l_0}) - \varphi_2(z^1, \ldots, z^{l_0})$$

is convex upward.

Consider the set

$$A^{II} = \{(z^1, \ldots, z^{l_0}): \varphi_3(z^1, \ldots, z^{l_0}) > a\}. \tag{A.15}$$

This set is convex and is located in E_{l_0}.

For the set A^{II} one of two inequalities is fulfilled: either

$$V(A^{II}) > C^{n-1}_{l_0} a^{l_0 - n + 1}, \tag{A.16}$$

or

$$V(A^{II}) \leq C^{n-1}_{l_0} a^{l_0 - n + 1}. \tag{A.17}$$

Assume that (A.16) is fulfilled. Then by the induction assumption there exists a coordinate space E_{n-1} of the space E_l such that the projection A^{III} of the set A^{II} on it contains an $(n-1)$-dimensional quasicube Ω_{n-1} with an edge a. Consider now the n-dimensional coordinate subspace E_n formed by E_{n-1} and the coordinate z^n. Furthermore let A^{IV} be the projection of the set A on the subspace E_n. For a given point $(z^1_*, \ldots, z^{n-1}_*) \in A^{III}$ consider the set $d = d(z^1_*, \ldots, z^{n-1}_*)$ of values of z such that $(z^1_*, \ldots, z^{n-1}_*, z) \in A^{IV}$.

It is easy to see that the set d contains an interval with end points

$$r_1(z^1, \ldots, z^{n-1}) = \sup'_{z \in A^{II}} \varphi_1(z^1, \ldots, z^{l_0}),$$

$$r_2(z^1, \ldots, z^{n-1}) = \inf'_{z \in A^{II}} \varphi_2(z^1, \ldots, z^{l_0}),$$

where sup′ and inf′ are taken over the points $z \in A^{II}$ which are projected onto a given point $(z^1_*, \ldots, z^{n-1}_*)$. Clearly, in view of (A.15), $r_1 - r_2 > a$. We now assign to each point $(z^1, \ldots, z^{n-1}) \in A^{III}$ a segment $c(z^1, \ldots, z^{n-1})$ of length a on the axis z^{l_0+1}:

$$[\tfrac{1}{2}(r_1(z^1, \ldots, z^{n-1}) + r_2(z^1, \ldots, z^{n-1})) - a/2,$$
$$\tfrac{1}{2}(r_1(z^1, \ldots, z^{n-1}) + r_2(z^1, \ldots, z^{n-1})) + a/2].$$

Clearly, $c(z^1, \ldots, z^{n-1}) \subset d(z^1, \ldots, z^{n-1})$.

Consider now the set $Q \subset E_n$ consisting of points $(z^1, \ldots, z^{n-1}, z^{l_0+1})$ such that

$$(z^1, \ldots, z^{n-1}) \in \Omega_{n-1}, \tag{A.18}$$

$$z^{l_0+1} \in c(z^1, \ldots, z^{n-1}). \tag{A.19}$$

This set is the required quasicube Ω_n. Indeed, in view of (A.18) and (A.19) the set Q satisfies the definition of an n-dimensional quasicube with an edge a. At the same time we have $Q \in A^{IV}$ by construction.

To prove the lemma it remains to consider the case when the inequality (A.17) is fulfilled, i.e.,

$$V(A^{II}) \leq C_{l_0}^{n-1} a^{l_0-n+1}.$$

Then

$$\begin{aligned} V(A) &= \int_{A^I} \varphi_3(z^1, \ldots, z^{l_0})\, dz^1 \cdots dz^{l_0} \\ &= \int_{A^I - A^{II}} \varphi_3(z^1, \ldots, z^{l_0})\, dz^1 \cdots dz^{l_0} \\ &\quad + \int_{A^{II}} \varphi_3(z^1, \ldots, z^{l_0})\, dz^1 \cdots dz^{l_0} \\ &\leq aV(A^I) + V(A^{II}), \end{aligned}$$

and in view of (A.14) and (A.17) we obtain

$$V(A) \leq C_{l_0}^n a^{l_0-n+1} + C_{l_0}^{n-1} a^{l_0-n+1} = C_{l_0+1}^n a^{(l_0+1)-n},$$

which contradicts the lemma's condition. □

§A3 ε-extension of a Set

Let A be a convex bounded set in E_l. We assign to each point $z \in A$ an open cube $\Omega(z)$ with the center at z and the edge ε oriented along the coordinate axes.

Consider the set

$$A_\varepsilon = \bigcup_{z \in A} \Omega(z),$$

along with the set A, which we shall call an ε-extension of the set A. The set A_ε is the set of points $y = (y^1, \ldots, y^l)$ for each of which there exists a point $z \in A$ such that

$$\rho(z, y) < \frac{\varepsilon}{2}.$$

It is easy to show that an ε-extension A_ε of the convex set A is convex.

Now choose a minimal proper ε-net on the set A. Let the minimal number of elements of a proper ε-net of the set A be $N_0(\varepsilon, A)$. Denote by $V(A_\varepsilon)$ the volume of the set A_ε.

Lemma A.2. *The inequality*

$$N_0(1.5\varepsilon, A)\varepsilon^l \leq V(A_\varepsilon) \tag{A.20}$$

is valid.

PROOF. Let T be a proper $\varepsilon/2$-net of the set A. Select a subset $\hat{T}$ of the set T according to the following rule:

(1) The first point $\hat{z}_1$ of the set $\hat{T}$ is an arbitrary point of T.
(2) Let m distinct points $\hat{z}_1, \ldots, \hat{z}_m$ be chosen. An arbitrary point of $z \in T$ such that

$$\min_{1 \leq i \leq m} \rho(z, \hat{z}_i) \geq \varepsilon$$

is selected as an $(m + 1)$th point of $\hat{T}$.
(3) If there is no such point or if T has been exhausted, then the construction is completed.

The set $\hat{T}$ constructed in the manner described above is a 1.5ε-net in A. Indeed, for any $z \in A$, there exists $t \in T$ such that $\rho(z, t) < \varepsilon/2$. For such a t there exists $\hat{z} \in \hat{T}$ such that $\rho(\hat{z}, t) < \varepsilon$. Consequently, $\rho(z, \hat{z}) < 1.5\varepsilon$ and the number of elements in T is at least $N_0(1.5\varepsilon, A)$.

Furthermore, the union of open cubes with edge ε and centers at the points of $\hat{T}$ is included in A_ε. At the same time cubes with centers at different points do not intersect. (Otherwise, there would exist $\hat{z} \in \Omega(z_1)$ and $\hat{z} \in \Omega(z_2)$, $z_1, z_2 \in \hat{T}$, and hence $\rho(z_1, \hat{z}) < \varepsilon/2$ and $\rho(z_2, z) < \varepsilon/2$, whence $\rho(z_1, z_2) < \varepsilon$ and $z_1 = z_2$.) Consequently

$$V(A_\varepsilon) \geq N_0(1.5\varepsilon, A)\varepsilon^l.$$

The lemma is proved. □

Lemma A.3. *Let a convex set A belong to the unit cube in E_l, and A_ε be its ε-extension $(0 < \varepsilon \leq 1)$; and for some $\gamma > \ln(1 + \varepsilon)$ let the inequality*

$$N_0(1.5\varepsilon, A) > e^{\gamma l}$$

be fulfilled. Then there exist $t(\varepsilon, \gamma)$ and $a(\varepsilon, \gamma)$ such that—provided $n = [t_0 l] > 0$—one can find a coordinate subspace of dimension $n = [t_0 l]$ such that a projection of A_ε on this space contains an n-dimensional quasicube with an edge a.

PROOF. In view of Lemmas A.1 and A.2 and the condition (A.20), which is valid for this lemma, in order that there exist an n-dimensional coordinate subspace such that the projection of A_ε on this space contains an n-dimensional quasicube with an edge a, it is sufficient that

$$C_l^n b^{l-n} < e^{\gamma l}\varepsilon^l(1 + \varepsilon)^{-l},$$

where $b = a/(1 + \varepsilon)$.

In turn it follows from Stirling's formula that for this purpose it is sufficient that

$$b^{l-n}\frac{l^n e^n}{n^n} < e^{\gamma_1 l}\varepsilon^l,$$

where $\gamma_1 = \gamma \ln(1+\varepsilon)$. Setting $t = n/l$ and taking $0 < t < \frac{1}{3}$, we obtain

$$-\frac{t(\ln t - 1)}{1-t} + \ln b < \frac{\ln \varepsilon + \gamma_1}{1-t},$$

using an equivalent transformation.

Under the stipulated restrictions this equality will be fulfilled if the inequality

$$-\tfrac{3}{2}t(\ln t - 1) + \ln b < (1+2t)\ln \varepsilon + \tfrac{2}{3}\gamma_1 \tag{A.21}$$

is satisfied. Now choose $t_0(\gamma, \varepsilon)$ such that the conditions

$$0 < t_0(\varepsilon, \gamma) \le \tfrac{1}{3},$$
$$-\tfrac{3}{2}t_0(\ln t_0 - 1) < \gamma_1/6,$$
$$-2t_0 \ln \varepsilon < \gamma_1/6$$

will be satisfied. This can always be achieved, since by assumption $\gamma_1 > 0$. Clearly for $0 < t \le t_0$ these conditions are also fulfilled and in this case (A.21) will be fulfilled for

$$\ln b = \ln \varepsilon + \frac{\gamma_1}{3},$$

or

$$a = (1+\varepsilon)\varepsilon \exp\left\{\frac{\gamma - \ln(1-\varepsilon)}{3}\right\}. \tag{A.22}$$

The lemma is thus proved. □

§A4 An Auxiliary Lemma

Now consider a class of functions $\Phi = F(x, \alpha)$ parametrized by means of $\alpha \in \Lambda$ defined on X. We shall assume that the class is convex in the sense that if

$$F(x, \alpha_1), \ldots, F(x, \alpha_r) \subset \Phi, \tag{A.23}$$

then

$$\sum_{i=1}^{r} \tau_i F(x, \alpha_i) \subset \Phi, \qquad \sum_{i=1}^{r} \tau_i = 1, \qquad \tau_i \ge 0.$$

Now define two sequences: the sequence

$$x_1, \ldots, x_l, \qquad x_i \in X,$$

and a random independent numerical sequence

$$y_1, \ldots, y_l, \tag{A.24}$$

which has the property

$$y_i = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}. \end{cases}$$

Using these sequences, we define the quantity

$$Q(\Phi) = M_y \sup_{F(x,\alpha)\in\Phi} \frac{1}{l} \left| \sum_{i=1}^{l} F(x_i, \alpha) y_i \right|.$$

(The expectation is taken over the random sequences (A.24).)

In Section A.1 we denoted by A the set of l-dimensional vectors z with coordinates $z^i = F(x_i, \alpha)$, $i = 1, \ldots, l$, for all possible $\alpha \in \Lambda$. Clearly A belongs to the unit l-dimensional cube in E_l and is convex.

We rewrite the function $Q(\Phi)$ in the form

$$Q(\Phi) = M_y \sup_{z\in A} \left| \frac{1}{l} \sum_{i=1}^{l} z^i y_i \right|.$$

The following lemma is valid.

Lemma A.4. *If for $\varepsilon > 0$ the inequality*

$$N_0(1.5\varepsilon, A) > e^{\gamma l}, \qquad \gamma > \ln(1 + \varepsilon),$$

is fulfilled for the set A, then the inequality

$$Q(\Phi) \geq \varepsilon \left(\exp\left\{ \frac{\gamma - \ln(1+\varepsilon)}{3} \right\} - 1 \right) \left(\frac{t}{2} - \frac{1}{2l} \right)$$

is valid, where $t > 0$ does not depend on l.

PROOF. As was shown in the preceding section, if the conditions of the lemma are fulfilled, there exist $t(\varepsilon, \gamma)$ and $a(\varepsilon, \gamma)$ such that there exists a coordinate subspace of dimension $n = [tl]$ with the property that a projection of the set A_ε on this subspace contains an n-dimensional quasicube with edge a. We have assumed here without loss of generality that this subspace forms the first n coordinates and the corresponding n-dimensional subspace forms a canonical subspace of this quasicube.

We define the vertices of the quasicube using the following iterative rule:

(1) The vertices of the one-dimensional cube are the end points of the segment c and $c + a$.

(2) To define vertices of an n-dimensional quasicube in an n-dimensional canonical space, we proceed as follows. Let the vertices of an $(n-1)$-dimensional quasicube be determined. Assign the segment

$$\left[\varphi^{n-1}(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}) - \frac{a}{2}, \varphi^{n-1}(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}) + \frac{a}{2}\right]$$

to each such vertex $(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1})$ (k is the number of the vertex), where

$$\varphi^{n-1}(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}) = \tfrac{1}{2}(\varphi_1(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}) + \varphi_2(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1})),$$

$$\varphi_1(\hat{z}^1, \ldots, \hat{z}^{n-1}) = \max_{\hat{z}^n} \{\hat{z}^n : (\hat{z}^1, \ldots, \hat{z}^{n-1}, \hat{z}^n) \in \Omega_n\},$$

$$\varphi_2(\hat{z}^1, \ldots, \hat{z}^{n-1}) = \min_{\hat{z}^n} \{\hat{z}^n : \hat{z}^1, \ldots, \hat{z}^{n-1}, \hat{z}^n) \in \Omega_n\},$$

and Ω_n is an n-dimensional quasicube.

This segment is formed by the intersection of the line $(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}, z^n)$ with the quasicube. The endpoints of the segment form the vertices of the quasicube. Thus if

$$(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}) \in E_{n-1}$$

is the kth vertex of an $(n-1)$-dimensional quasicube, then

$$\left(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}, \varphi^{n-1}(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}) + \frac{a}{2}\right),$$

$$\left(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}, \varphi^{n-1}(\hat{z}_k^1, \ldots, \hat{z}_k^{n-1}) - \frac{a}{2}\right)$$

are correspondingly the $(2k-1)$th and the $2k$th vertices of the n-dimensional quasicube.

Now we assign to an arbitrary sequence

$$y_1, \ldots, y_n \qquad \left(y_i = \begin{Bmatrix} +1 \\ -1 \end{Bmatrix}\right)$$

a vertex $\hat{z}_*$ of a quasicube defined as follows:

$$\hat{z}_*^1 = \left(c + \frac{a}{2}\right) + \frac{a}{2} y_1,$$

$$\hat{z}_*^j = \varphi^{j-1}(\hat{z}_*^1, \ldots, \hat{z}_*^{j-1}) + \frac{a}{2} y_j, \qquad j = 2, \ldots, n.$$

In turn, to each vertex $\hat{z}_*$ of a quasicube in E_n we assign a point $z_* = (z_*^1, \ldots, z_*^l) \in A$ such that the distance between the projection $(z_*^1, \ldots, z_*^n)$ of this point in E_n and the vertex $\hat{z}_*$ is at most $\varepsilon/2$, i.e.,

$$|z_*^j - \hat{z}_*^j| < \frac{\varepsilon}{2}, \qquad j = 1, 2, \ldots, n.$$

This is possible because $z_* \in \Pr A_\varepsilon$ on E_n.

Thus we introduce two functions

$$\hat{z}_* = \hat{z}_*(y_1, \ldots, y_n),$$
$$z_* = z_*(\hat{z}_*^1, \ldots, \hat{z}_*^n).$$

We shall denote the difference $z_*^j - \hat{z}_*^j$ by δ_j $(j = 1, \ldots, n)$ $(|\delta_j| \leq \varepsilon/2)$ and bound the quantity

$$\begin{aligned} Q(\Phi) &= M \sup_{z \in A} \frac{1}{l} \left| \sum_{i=1}^{l} z^i y_i \right| \\ &\geq \frac{1}{l} M \sum_{i=1}^{l} z_*^i y_i \\ &= \frac{1}{l} \sum_{i=1}^{n} M y_i (\hat{z}_*^i + \delta_i) + \frac{1}{l} \sum_{i=n+1}^{l} M y_i z_*^i. \end{aligned}$$

Observe that the second summand in the sum is zero, since every term of the sum is a product of two independent random variables y_i and z_*^i, $i > n$, one of which (y_i) has zero mean.

We shall bound the first summand. For this purpose consider the first term in the first summand:

$$\begin{aligned} &\frac{1}{l} M \left[y_1 \left(c + \frac{a}{2} + \frac{a}{2} y_1 + \delta_1 \right) \right] \\ &\quad = \frac{1}{l} \left[\frac{a}{2} + M y_1 \delta_1 \right] \\ &\quad \geq \frac{1}{2l} (\alpha - \varepsilon). \end{aligned}$$

To bound the kth term

$$I_k = \frac{1}{l} M \left[y_k (\varphi^{k-1}(\hat{z}_*^1, \ldots, \hat{z}_*^{k-1}) + \frac{a}{2} y_k + \delta_k) \right],$$

we observe that the vertex $(\hat{z}_*^1, \ldots, \hat{z}_*^{k-1})$ was chosen in such a manner that it would not depend on y_k but only on $y_1, \ldots, y_{k-1}$. Therefore

$$I_k = \frac{1}{l} \left[\frac{a}{2} + M y_k \delta_k \right] \geq \frac{1}{2l} (a - \varepsilon).$$

Thus we obtain

$$Q(\Phi) > M \sup_{z_* \in A} \frac{1}{l} \sum_{i=1}^{l} z_*^i y_i \geq \frac{n}{2l} (a - \varepsilon) > (a - \varepsilon) \left(\frac{t}{2} - \frac{1}{2l} \right).$$

Choosing the quantity a in accordance with (A.22), we arrive at

$$Q(\Phi) > \varepsilon \left(\exp\left\{ \frac{\gamma - \ln(1 + \varepsilon)}{3} \right\} - 1 \right) \left(\frac{t}{2} - \frac{1}{2l} \right).$$

The lemma is thus proved. □

§A5 Necessary and Sufficient Conditions for Uniform Convergence: The Proof of Necessity

Theorem A.2. *For the uniform convergence of the means to their mathematical expectations over a uniformly bounded class of functions* $F(x, \alpha)$, $\alpha \in \Lambda$, *it is necessary and sufficient that for any* $\varepsilon > 0$ *the equality*

$$\lim_{l \to \infty} \frac{H^{\Lambda}(\varepsilon, l)}{l} = 0 \tag{A.25}$$

be satisfied.

To prove the necessity we can assume without loss of generality that the class $F(x, \alpha)$ is convex in the sense of (A.23), since from the uniform convergence of the means to their mathematical expectations for an arbitrary class follows the same convergence for its convex closure, and the condition (A.25) for a convex closure implies the same for the initial class of functions.

PROOF OF NECESSITY. Assume the contrary. For some $\varepsilon_0 > 0$ let the equality

$$\lim_{l \to \infty} \frac{H^{\Lambda}(\varepsilon_0, l)}{l} = c(\varepsilon_0) > 0 \tag{A.26}$$

be fulfilled, and at the same time let uniform convergence hold, i.e., for all ε let the relationship

$$\lim_{l \to \infty} P\left\{\sup_{\alpha \in \Lambda} \left| MF(x, \alpha) - \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha) \right| > \varepsilon \right\} = 0 \tag{A.27}$$

be satisfied. This will lead to a contradiction.

Since the functions $MF(x, \alpha)$, $(1/l) \sum_{i=1}^{l} F(x_i, \alpha)$, $\alpha \in \Lambda$, are uniformly bounded by 1, it follows from (A.27) that

$$\lim_{l \to \infty} M\left\{\sup_{\alpha \in \Lambda} \left| MF(x, \alpha) - \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha) \right| \right\} = 0.$$

This implies that if $l_1 \to \infty$ and $l - l_1 \to \infty$, then the equality

$$\lim_{l_1, l \to \infty} M\left\{\sup_{\alpha \in \Lambda} \left| \frac{1}{l_1} \sum_{i=1}^{l_1} F(x_i, \alpha) - \frac{1}{l - l_1} \sum_{i=l_1+1}^{l} F(x_i, \alpha) \right| \right\} = 0 \tag{A.28}$$

is fulfilled.

Consider the expression

$$I(x_1, \ldots, x_l) = \sum_{n=0}^{l} \sup_{\alpha \in \Lambda} \left[\frac{C_l^n}{2^l} \frac{1}{l} \left| \sum_{i=1}^{n} F(x_i, \alpha) - \sum_{i=n+1}^{l} F(x_i, \alpha) \right| \right].$$

We subdivide the summation with respect to n into two "regions":

$$\text{I:} \quad \left| n - \frac{l}{2} \right| < l^{2/3},$$

$$\text{II:} \quad \left| n - \frac{l}{2} \right| \geq l^{2/3}.$$

Then taking into account that

$$\frac{1}{l} \left| \sum_{i=1}^{n} F(x_i, \alpha) - \sum_{i=n+1}^{l} F(x_i, \alpha) \right| \leq 1,$$

we obtain

$$\begin{aligned} I(x_1, \ldots, x_l) \leq & \sum_{n \in \text{II}} \frac{C_l^n}{2^l} \\ & + \sum_{n \in \text{I}} \frac{C_l^n}{2^l} \sup_{\alpha \in \Lambda} \left| \frac{1}{n} \left(\frac{n}{l} \sum_{i=1}^{n} F(x_i, \alpha) \right) \right. \\ & \left. - \frac{l-n}{l} \left(\frac{1}{l-n} \sum_{i=n+1}^{l} F(x_i, \alpha) \right) \right|. \end{aligned}$$

Note that in region I ($\frac{1}{2} - 1/l^{1/3} < n/l < \frac{1}{2} + 1/l^{1/3}$),

$$\sum_{n \in \text{I}} \frac{C_l^n}{2^l} \xrightarrow[l \to \infty]{} 1,$$

while in region II

$$\sum_{n \in \text{II}} \frac{C_l^n}{2^l} \xrightarrow[l \to \infty]{} 0. \tag{A.29}$$

Furthermore

$$\begin{aligned} \lim_{l \to \infty} MI(x_1, \ldots, x_l) \leq \lim_{l \to \infty} & \left(\sum_{n \in \text{II}} \frac{C_l^n}{2^l} \right. \\ & \left. + \frac{1}{2} \max_{n \in \text{I}} M \sup_{\alpha \in \Lambda} \left| \frac{1}{n} \sum_{i=1}^{n} F(x_i, \alpha) - \frac{1}{l-n} \sum_{i=n+1}^{l} F(x_i, \alpha) \right| \sum_{n \in \text{I}} \frac{C_l^n}{2^l} \right). \end{aligned}$$

It follows from (A.28) that

$$\max_{n \in \text{I}} M \sup_{\alpha \in \Lambda} \left| \frac{1}{n} \sum_{i=1}^{n} F(x_i, \alpha) - \frac{1}{l-n} \sum_{i=n+1}^{l} F(x_i, \alpha) \right| \xrightarrow[l \to \infty]{} 0.$$

Thus taking (A.29) into account we have

$$\lim_{l \to \infty} MI(x_1, \ldots, x_l) = 0. \tag{A.30}$$

On the other hand

$$MI(x_1, \ldots, x_l) = M \frac{1}{l!} \sum_{k=1}^{l!} I(T_k\{x_1, \ldots, x_l\}),$$

where T_k ($k = 1, \ldots, l!$) are all the permutations of the sequence. We transform the right-hand side:

$$M\frac{1}{l!}\sum_{k=1}^{l!} I(T_k\{x_1, \ldots, x_l\})$$

$$= M\frac{1}{l!}\sum_{k=1}^{l!}\sum_{n=0}^{l}\sup_{\alpha\in\Lambda}\left[\frac{C_l^n}{2^l}\frac{1}{l}\left|\sum_{i=1}^{n} F(x_{j(i,k)}, \alpha) - \sum_{i=n+1}^{l} F(x_{j(i,k)}, \alpha)\right|\right]$$

$$= M\sum_{n=0}^{l}\frac{1}{C_l^n}\sum_{y_1,\ldots,y_l}\sup_{\alpha\in\Lambda}\frac{C_l^n}{2^l}\frac{1}{l}\left|\sum_{i=1}^{l} y_i F(x_i, \alpha)\right|.$$

(Here $j(i, k)$ is the index obtained when the permutation T_k acts on i.) In the last expression the summation is carried out over all the sequences

$$y_1, \ldots, y_l \qquad \left(y_i = \begin{cases} +1 \\ -1 \end{cases}\right)$$

which have n positive values.

Furthermore we obtain

$$MI(x_1, \ldots, x_l) = M\frac{1}{2^l}\left\{\sum_{y_1,\ldots,y_l}\sup_{\alpha\in\Lambda}\frac{1}{l}\left|\sum_{i=1}^{l} y_i F(x_i, \alpha)\right|\right\}. \tag{A.31}$$

In (A.31) the summation is carried over all sequences

$$y_1, \ldots, y_l.$$

Choose for $\varepsilon_0 > 0$ a number such that

$$\lim_{l\to\infty}\frac{H^{\Lambda}(\varepsilon_0 l)}{l} = c(\varepsilon) > 0.$$

Since $c(\varepsilon)$ is nondecreasing as ε decreases, one can choose ε in such a manner that

$$0 < 1.5\varepsilon \le \varepsilon_0, \qquad \ln(1 + \varepsilon) < \frac{c(\varepsilon) - \ln 2}{2}, \qquad c(1.5\varepsilon) \ge c(\varepsilon_0)$$

will be fulfilled. Then in view of (A.9) the probability that the inequality

$$N_0^{\Lambda}(x_1, \ldots, x_l, 1.5\varepsilon) > \exp\left\{\frac{c(\varepsilon_0)\ln 2}{2}\right\} \tag{A.32}$$

is fulfilled approaches 1.

According to Lemma A.4, if (A.32) is satisfied, the expression appearing in the braces in (A.31) exceeds

$$\varepsilon\left(\frac{t}{2} - \frac{1}{2l}\right)\left(\exp\left\{\frac{\gamma}{3}\right\} - 1\right),$$

where $\gamma = \frac{1}{2}c(\varepsilon_0)\ln 2 - \ln(1+\varepsilon)$, and $t(\varepsilon, \gamma)$ does not depend on l. From this we conclude that

$$\lim_{l\to\infty} I(x_1, \ldots, x_l) > \lim_{l\to\infty} \varepsilon\left(\frac{t}{2} - \frac{1}{2l}\right)(e^{\gamma/3} - 1) > 0.$$

This inequality contradicts the assertion (A.30), and the contradiction obtained proves the first part of the theorem. □

§A6 Necessary and Sufficient Conditions for Uniform Convergence: The Proof of Sufficiency

The following lemma is valid.

Lemma A.5. *If for any $\varepsilon > 0$ the relation*

$$P\left\{\sup_{\alpha\in\Lambda}\left|\frac{1}{l}\sum_{i=1}^{l} F(x_i, \alpha) - \frac{1}{l}\sum_{i=l+1}^{2l} F(x_i, \alpha)\right| > \varepsilon\right\} \underset{l\to\infty}{\longrightarrow} 0 \qquad \text{(A.33)}$$

is valid, then for any ε the relation

$$P\left\{\sup_{\alpha\in\Lambda}\left|\frac{1}{l}\sum_{i=1}^{l} F(x_i, \alpha) - MF(x, \alpha)\right| > \varepsilon\right\} \underset{l\to\infty}{\longrightarrow} 0$$

also holds.

PROOF. Assume the contrary. For $\varepsilon_0 > 0$ let

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left|\frac{1}{l}\sum_{i=1}^{l} F(x_i, \alpha) - MF(x, \alpha)\right| > \varepsilon_0\right\} \neq 0.$$

Denote by R_l the event

$$\sup_{\alpha\in\Lambda}\left|\frac{1}{l}\sum_{i=1}^{l} F(x_i, \alpha) - MF(x, \alpha)\right| > \varepsilon_0\Bigg\}.$$

Then for l sufficiently large the inequality

$$P\{R_l\} > \eta > 0$$

is fulfilled. Denote

$$\frac{1}{l}\left|\sum_{i=1}^{l} F(x_i, \alpha) - \sum_{i=l+1}^{2l} F(x_i, \alpha)\right| = S(x_1, \ldots, x_l, \alpha)$$

and consider the quantity

$$P_{2l} = P\left\{\sup_{\alpha\in\Lambda} S(x_1, \ldots, x_{2l}, \alpha) > \frac{\varepsilon_0}{3}\right\}$$

$$= \int_{x_1,\ldots,x_{2l}} \cdots \int \theta\left[\sup_{\alpha\in\Lambda} S(x_1, \ldots, x_{2l,\alpha}) - \frac{\varepsilon_0}{3}\right] dP(x_1)\cdots dP(x_{2l}).$$

Next the inequality

$$P_{2l} \geq \int_{R_l} \left\{ \int \cdots \int_{x_1, \ldots, x_{2l}} \theta\left[\sup_{\alpha \in \Lambda} S(x_1, \ldots, x_{2l}, \alpha) - \frac{\varepsilon_0}{3} \right] dP(x_1) \cdots dP(x_{2l}).$$

is valid. To each point $x_1, \ldots, x_l$ belonging to R_l we assign the value $\alpha^*(x_1, \ldots, x_l)$ such that

$$\left| \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha^*) - MF(x, \alpha^*) \right| > \frac{\varepsilon_0}{3}.$$

Denote by $\bar{R}_l$ the event in $X_l = (x_{l+1}, \ldots, x_{2l})$ such that

$$\frac{1}{l} \left| \sum_{i=l+1}^{2l} F(x_i, \alpha^*) - MF(x, \alpha^*) \right| \leq \frac{\varepsilon_0}{3}.$$

Since the function $F(x, \alpha)$ is uniformly bounded, it follows that

$$P(\bar{R}_l) \underset{l \to \infty}{\longrightarrow} 1.$$

Furthermore

$$P_{2l} \geq \int_{R_l} \left\{ \int_{R_l} \theta\left[S(x_1, \ldots, x_{2l}; \alpha^*(x_1, \ldots, x_l)) - \frac{\varepsilon_0}{3} \right] \times dP(x_{l+1}) \cdots dP(x_{2l}) \right\} dP(x_1) \cdots dP(x_l).$$

However if, $x_1, \ldots, x_l \in R_l$ and $x_{l+1}, \ldots, x_{2l} \in \bar{R}_l$, then the integrand equals 1. Choosing l so large that $P(\bar{R}_l) > \frac{1}{2}$, we obtain

$$P_{2l} > \frac{1}{2} \int_{R_l} dP(x_1) \cdots dP(x_l) = \tfrac{1}{2} P(R_l),$$

and hence $\lim_{l \to \infty} P_l \neq 0$, which contradicts the lemma's assumption. □

PROOF OF SUFFICIENCY. We shall prove that under the conditions of the theorem

$$P\left\{ \sup_{\alpha \in \Lambda} S(x_1, \ldots, x_{2l}; \alpha) > \varepsilon \right\} \underset{l \to \infty}{\longrightarrow} 0.$$

In view of Lemma A.5 it follows from the condition (A.33) that the assertion of the theorem is valid:

$$P\left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha) - MF(x, \alpha) \right| > \varepsilon \right\} \underset{l \to \infty}{\longrightarrow} 0.$$

We shall now verify (A.33).

For this purpose observe that since the measure is by definition symmetric, the equality

$$P\left\{\sup_{\alpha\in\Lambda} S(x_1,\ldots,x_{2l},\alpha) > \varepsilon\right\}$$

$$= \frac{1}{(2l)!}\sum_{j=1}^{(2l)!} P\left\{\sup_{\alpha\in\Lambda} S(T_j\{x_1,\ldots,x_{2l}\},\alpha) > \varepsilon\right\}$$

$$= \int\left[\frac{1}{(2l)!}\sum_{j=1}^{(2l)!}\theta\left[\sup_{\alpha\in\Lambda} S(T_j\{x_1,\ldots,x_{2l}\},\alpha) - \varepsilon\right]\right]$$

$$\times\, dP(x_1)\cdots dP(x_{2l}) \tag{A.34}$$

is valid; here $T_j, j = 1,\ldots,(2l)!$, are all the permutations of the indices, and $T_j(x_1,\ldots,x_{2l})$ is a sequence of arguments obtained from the sequence $x_1,\ldots,x_{2l}$ when the permutation T_j is applied.

Now consider the integrand in (A.34):

$$K = \frac{1}{(2l)!}\sum_{j=1}^{(2l)!}\theta\left(\sup_{\alpha\in\Lambda} S(T_j\{x_1,\ldots,x_{2l}\},\alpha) - \varepsilon\right).$$

Let A be the set of points in E_{2l} with coordinates $z^i = F(x_i,\alpha)$, $i = 1,\ldots,2l$, for all $\alpha\in\Lambda$.

Let $z(1),\ldots,z(N_0)$ be the minimal proper ε-net in A, and $\alpha(1),\ldots,\alpha(N_0)$ be the values of α such that

$$z^i(k) = F(x_i,\alpha(k)), \qquad i = 1,\ldots,2l, \quad k = 1,\ldots,N_0.$$

We show that if the inequality

$$\max_{1\le k\le N_0} S(x_1,\ldots,x_{2l};\alpha(k)) < \frac{\varepsilon}{3}$$

is fulfilled, then the inequality

$$\sup_{\alpha\in\Lambda} S(x_1,\ldots,x_{2l},\alpha) < \varepsilon$$

is also valid.

Indeed, for any α there exists $\alpha(k)$ such that

$$|F(x_i,\alpha) - F(x_i,\alpha(k)| < \frac{\varepsilon}{3}, \qquad i = 1,2,\ldots,2l.$$

Therefore

$$\left|\frac{1}{l}\sum_{i=1}^{l}F(x_i,\alpha)-\frac{1}{l}\sum_{i=l+1}^{2l}F(x_i,\alpha)\right|$$

$$=\left|\frac{1}{l}\left(\sum_{i=1}^{l}F(x_i,\alpha)-\sum_{i=1}^{i}F(x_i,\alpha(k))\right)\right.$$

$$-\frac{1}{l}\left(\sum_{i=l+1}^{2l}F(x_i,\alpha)-\sum_{i=l+1}^{2l}F(x_i,\alpha(k))\right)$$

$$\left.+\frac{1}{l}\left(\sum_{i=1}^{l}F(x_i,\alpha(k))-\sum_{i=l+1}^{2l}F(x_i,\alpha(k))\right)\right|$$

$$\leq 2\frac{\varepsilon}{3}+\frac{1}{l}\left|\sum_{i=1}^{l}F(x_i,\alpha(k))-\sum_{i=l+1}^{2l}F(x_i,\alpha(k))\right|<\varepsilon.$$

Analogous bounds are valid for $S(T_j\{x_1,\ldots,x_{2l}\},\alpha)$. Therefore

$$K=\frac{1}{(2l)!}\sum_{j=1}^{(2l)!}\theta\left[\max_k S(T_j\{x_1,\ldots,x_{2l}\},\alpha(k))-\frac{\varepsilon}{3}\right]$$

$$\leq\frac{1}{(2l)!}\sum_{j=1}^{(2l)!}\sum_{k=1}^{N_0}\theta[S(T_j\{x_1,\ldots,x_{2l}\},\alpha(k))-\frac{\varepsilon}{3}]$$

$$=\sum_{k=1}^{N_0}\left\{\frac{1}{(2l)!}\sum_{j=1}^{(2l)!}\theta\left[S(T_j\{x_1,\ldots,x_{2l}\},\alpha(k))-\frac{\varepsilon}{3}\right]\right\}.$$

We now bound the expression in the braces:

$$R_1=\frac{1}{(2l)!}\sum_{j=1}^{(2l)!}\theta\left(\left|\frac{1}{l}\sum_{i=1}^{l}F(x_{T_j(i)},\alpha(k)-\frac{1}{l}\sum_{i=l+1}^{2l}F(x_{T_j(i)},\alpha(k))\right|-\frac{\varepsilon}{3}\right).$$

Here $T_j(i)$ is the index obtained when the permutation T_j acts on i.

We arrange the values

$$F(x_{i_1},\alpha(k))\leq F(x_{i_2},\alpha(k)\leq\cdots\leq F(x_{i_{2l}},\alpha(k))$$

in the order of their magnitudes and denote $z^p=F(x_{i_p},\alpha(k))$.

Next we use the notation

$$\Delta_1=z^1,\qquad \Delta_p=z^p-z^{p-1},$$

$$\delta_{ip}=\begin{cases}1 & \text{for } F(x_i,\alpha(k))\leq z^p,\\ 0 & \text{for } F(x_i,\alpha(k))>z^p,\end{cases}$$

$$r_i^j=\begin{cases}1 & \text{for } T_j^{-1}(i)\leq l,\\ 0 & \text{for } T_j^{-1}(i)>l,\end{cases}$$

where $T_j^{-1}(i)$ is the index which is mapped into i by the permutation T_j. Then

$$\frac{1}{l}\left|\sum_{i=1}^{l} F(x_{T_j(i)}, \alpha(k)) - \sum_{i=l+1}^{2l} F(x_{T_j(i)}, \alpha(k))\right|$$
$$= \frac{1}{l}\left|\sum_p \Delta_p \sum_{i=1}^{2l} \delta_{ip} r_i^j - \sum_p \Delta_p \sum_{i=1}^{2l} \delta_{ip}(1 - r_i^j)\right|$$
$$= \sum_p \Delta_p \left|\frac{1}{l}\sum_{i=1}^{2l} \delta_{ip}(2r_i^j - 1)\right|.$$

Furthermore, if the inequality

$$\max_p \frac{1}{l}\left|\sum_{i=1}^{2l} \delta_{ip}(2r_i^j - 1)\right| < \frac{\varepsilon}{3} \tag{A.35}$$

is fulfilled, then the inequality

$$\sum_p \Delta_p \frac{1}{l}\left|\sum_{i=1}^{2l} \delta_{ip}(2r_i^j - 1)\right| < \frac{\varepsilon}{3}\sum_p \Delta_p \leq \frac{\varepsilon}{3} \tag{A.36}$$

is also valid. The condition (A.35) is equivalent to the following

$$\max_p \theta\left[\frac{1}{l}\left|\sum_{i=1}^{2l} \delta_{ip}(2r_i^j - 1)\right| - \frac{\varepsilon}{3}\right] = 0.$$

Thus we obtain

$$R_1 < \frac{1}{(2l)!}\sum_{j=1}^{(2l)!} \max_p \theta\left[\frac{1}{l}\left|\sum_{i=1}^{2l} \delta_{ip}(2r_i^j - 1)\right| - \frac{\varepsilon}{3}\right]$$
$$\leq \sum_p \left\{\frac{1}{(2l)!}\sum_{j=1}^{(2l)!} \theta\left[\frac{1}{l}\left|\sum_{i=1}^{2l} \delta_{ip}(2r_i^j - 1)\right| - \frac{\varepsilon}{3}\right]\right\}. \tag{A.37}$$

Let there be $2l$ balls, of which $\sum_{i=1}^{2l} \delta_{ip} = m$ are black, in an urn model without replacement. We select l balls (without replacement). Then the expression in the braces of (A.37) is the probability that the number of black balls chosen from the urn will differ from the number of remaining black balls by at least $(\varepsilon/3)l$. This value equals

$$\Gamma = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

where k runs over all the values such that

$$\left|\frac{k}{l} - \frac{m-k}{l}\right| > \frac{\varepsilon}{3}.$$

In the Appendix to Chapter 6 the bound

$$\Gamma < 3\exp\left\{-\frac{\varepsilon^2 l}{9}\right\}$$

was derived. Thus

$$R_1 < \sum_{p=1}^{2l} 3 \exp\left\{-\frac{\varepsilon^2 l}{9}\right\} = 6l \exp\left\{-\frac{\varepsilon^2 l}{9}\right\}.$$

Returning to the bound, on K we obtain

$$K < 6lN_0\left(x_1, \ldots, x_{2l}, \frac{\varepsilon}{3}\right) \exp\left\{-\frac{e^2 l}{9}\right\}$$

Finally, for any $c > 0$ we have

$$\begin{aligned} P&\left\{\sup_{\alpha \in \Lambda} \frac{1}{l} \left| \sum_{i=1}^{l} F(x_i, \alpha) - \sum_{i=l+1}^{2l} F(x_i, \alpha) \right| > \varepsilon \right\} \\ &\leq \int_{\log_2 N_0^{\Lambda}(x_1, \ldots, x_{2l}; \varepsilon/\varepsilon) > cl} dP(x_1) \cdots dP(x_{2l}) \\ &\quad + \int_{\log_2 N_0^{\Lambda}(x_1, \ldots, x_{2l}; \varepsilon/3) \leq cl} K(x_1, \ldots, x_{2l})\, dP(x_1) \cdots dP(x_{2l}) \\ &\leq P\left\{\frac{\log_2 N_0^{\Lambda}(x_1, \ldots, x_{2l}; \varepsilon/3)}{l} > c\right\} \\ &\quad + 6l \exp\left\{-\frac{\varepsilon^2 l}{9} + cl\right\}. \end{aligned}$$

Setting $c < \varepsilon^2/10$, we obtain that the second term on the right-hand side approaches zero as l increases. In view of the condition of the theorem and the relation (A.8), the first term tends to zero. The theorem is proved. □

§A7 Corollaries

Theorem A.3. *For uniform convergence of means to their mathematical expectations it is necessary and sufficient that for any $\varepsilon > 0$ the equality*

$$\lim_{l \to \infty} \frac{1}{l} M \log V(A_\varepsilon) = \log \varepsilon$$

be fulfilled, where A_ε is the ε-extension of the set A.

PROOF. *Necessity.* Let $\varepsilon, \delta > 0$, $\delta < \varepsilon$ and T_0 be a minimal δ-net A with the number of elements $N_0^{\Lambda}(x_1 \ldots, x_l, \delta)$. We assign to each point in T_0 a cube with edge $\varepsilon + 2\delta$ and center at this point, oriented along the coordinate axes.

The union of these cubes contains A_ε, and hence

$$V(A_\varepsilon) < N_0^{\Lambda}(x_1, \ldots, x_l; \delta)(\varepsilon + 2\delta)^l;$$

whence we obtain

$$\lim_{l\to\infty} M\frac{1}{l}\log V(A_\varepsilon) \le \frac{H^\Lambda(\varepsilon, l)}{l} + \log(\varepsilon + 2\delta).$$

In view of the basic theorem,

$$M\frac{1}{l}\log V(A_\varepsilon) \le \log(\varepsilon + 2\delta).$$

Since $V(A_\varepsilon) > \varepsilon^l$ and δ is arbitrary, we arrive at the required assertion.

Sufficiency is obtained from the following considerations. Assume that the uniform convergence is not valid. Then for some $\varepsilon > 0$

$$\lim_{l\to\infty} M \log N_0^\Lambda(x_1, \ldots, x_l; 1.5\varepsilon) = \gamma > 0$$

whence in view of Lemma A.2

$$\lim_{l\to\infty} M\frac{\log V(A_\varepsilon)}{l} \ge \gamma + \log\varepsilon. \qquad \square$$

Lemma A.6. *If uniform convergence is valid in the class of functions* $F(x, \alpha)$, *it is then also valid in the class* $|F(x, \alpha)|$.

PROOF. The mapping

$$F(x, \alpha) \to |F(x, \alpha)|$$

does not increase the distance

$$\rho(\alpha_1, \alpha_2) = \max_{1\le i\le l} |F(x_i, \alpha_1) - F(x_i, \alpha_2)|.$$

Therefore

$$N_0^\Lambda(x_1, \ldots, x_l; \varepsilon) > \hat{N}_0^\Lambda(x_1, \ldots, x_l; \varepsilon),$$

where N_0^Λ and $\hat{N}_0^\Lambda$ are the minimal numbers of the elements in a ε-net in the sets A and A' respectively generated by the classes $F(x, \alpha)$ and $|F(x, \alpha)|$.

Consequently the condition

$$\lim_{l\to\infty} P\left\{\frac{\log N_0^\Lambda(x_1, \ldots, x_l; \varepsilon)}{l} > \delta\right\} = 0$$

implies

$$\lim_{l\to\infty} P\left\{\frac{\log \hat{N}_0^\Lambda(x_1, \ldots, x_l; \varepsilon)}{l} > \delta\right\} = 0$$

q.e.d. $\square$

Consider a two-parameter class of functions

$$f(x, \alpha_1, \alpha_2) = |F(x, \alpha_1) - F(x, \alpha_2)|, \qquad \alpha_1, \alpha_2 \in \Lambda,$$

along with the class of functions $F(x, \alpha)$, $\alpha \in \Lambda$.

Lemma A.7. *Uniform convergence in the class $F(x, \alpha)$ implies uniform convergence in $f(x, \alpha_1, \alpha_2)$.*

PROOF. Uniform convergence in $F(x, \alpha)$ clearly implies such a convergence in $F(x, \alpha_1) - F(x, \alpha_2)$. Indeed, the condition

$$\sup_{\alpha} \left| MF(x, \alpha) - \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha) \right| < \varepsilon$$

and the condition

$$\begin{aligned} &\left| MF(x, \alpha_1) - MF(x, \alpha_2) \right. \\ &\left. - \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha_1) + \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha_2) \right| \\ &\qquad \leq \left| MF(x, \alpha_1) - \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha_1) \right| \\ &\qquad + \left| MF(x, \alpha) - \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha_2) \right| \end{aligned}$$

imply that

$$\sup_{\alpha_1, \alpha_2} \left| M(F(x, \alpha_1) - F(x, \alpha_2)) - \frac{1}{l} \sum_{i=1}^{l} (F(x_i, \alpha_1) - F(x_i, \alpha_2)) \right| \leq 2\varepsilon.$$

Applying Corollary 2, we now obtain the required result. □

Denote by $L(x_1, \ldots, x_l, \varepsilon)$ the number of elements in the minimal ε-net of the set $A(x_1, \ldots, x_l)$ in the metric

$$\rho_1(z_1, z_2) = \frac{1}{l} \sum_{i=1}^{l} |z_1^i - z_2^i|.$$

Theorem A.4. *For a uniform convergence of means to mathematical expectations it is necessary and sufficient that a function $T(\varepsilon)$ exists such that*

$$\lim_{l \to \infty} P\{L(x_1, \ldots, x_l; \varepsilon) > T(\varepsilon)\} = 0.$$

PROOF. *Necessity.* The uniform convergence of $F(x, \alpha)$ implies the uniform convergence of the function $f(x, \alpha_1, \alpha_2)$, i.e.,

$$\sup_{\alpha_1, \alpha_2} \left| \frac{1}{l} \sum_{i=1}^{l} |F(x_i, \alpha_1) - F(x_i, \alpha_2)| - M|F(x, \alpha_1) - F(x, \alpha_2)| \right| \xrightarrow[l \to \infty]{p} 0. \quad \text{(A.38)}$$

Consequently for a finite l_0 and a given ε there exists a sequence $x_1^*, \ldots, x_l^*$ such that the left-hand side of (A.38) is smaller than ε. This means that the distance

$$\hat{\rho}_1(\alpha_1, \alpha_2) = \frac{1}{l_0} \sum_{i=1}^{l_0} |F(x_i^*, \alpha_1) - F(x_i, \alpha_2)| \tag{A.39}$$

approximates with precision ε the distance in the space of functions

$$\hat{\rho}_2(\alpha_1, \alpha_2) = \int |F(x, \alpha_1) - F(x, \alpha_2)| \, dP(x) \tag{A.40}$$

uniformly in α_1 and α_2. However, in the metric (A.39) there exists on the set Λ a finite ε-net S with the number of elements $L(x_1^*, \ldots, x_l^*; \varepsilon)$. The same net S forms a 2ε-net in the space Λ with the metric (A.40).

Next we utilize the uniform convergence of $\hat{\rho}(\alpha_1, \alpha_2)$ to $\hat{\rho}_2(\alpha_1, \alpha_2)$ and obtain that the same net S, with probability approaching 1 as $l \to \infty$, forms a 3ε-net on the set $A(x_1^*, \ldots, x_l^*)$. Setting $T(\varepsilon) = L(x_1^*, \ldots, x_l^*; \varepsilon)$, we obtain the assertion of the theorem.

The proof of sufficiency of the conditions of the theorem for uniform convergence is analogous to the proof of sufficiency for Theorem A.2. □

Chapter 8

The Method of Structural Minimization of Risk

§1 The Idea of the Method of Structural Risk Minimization

Up until now, when studying methods for estimation of dependences based on empirical data, the amount of data was of secondary importance: the principles which determined the selection of the desired dependence from a given set of possible dependences did not take into account directly the amount of available information.

Starting with this chapter, we shall consider methods of estimation which will allow us to obtain the best possible result (in a certain sense) for a given fixed amount of empirical data. Here it is essential to take into account the amount of available information, especially if the size of the sample

$$x_1, y_1; \ldots; x_l, y_l \tag{8.1}$$

is small. However, before proceeding to a discussion of methods for estimating dependences suited for a small sample, we shall clarify the meaning of a "small" sample.

Definition. We say that for purposes of estimating a function in a given class $F(x, \alpha)$ the sample size l is small if the ratio l/h is small (for example $l/h < 30$); here h is the capacity of the class of functions.

The quantity l/h determines the relative size of the sample (the sample size per unit capacity of the class).

Observe that the bounds on the expected risk obtained in Chapters 6 and 7 depend on the relative size of the samples rather than on their absolute

size. The basic result of Chapter 6 is that simultaneously for all indicator functions $F(x, \alpha)$ the inequality

$$P(\alpha) < \nu(\alpha) + \Omega_1\left(\frac{l}{h}, -\frac{\ln \eta}{h}\right) \tag{8.2}$$

is satisfied with probability $1 - \eta$, while the basic result of Chapter 7 is that simultaneously for the whole set of arbitrary functions $F(x, \alpha)$ the inequality

$$I(\alpha) < I_{\text{emp}}(\alpha)\Omega_2\left(\frac{l}{h}, \frac{-\ln \eta}{h}\right) \tag{8.3}$$

is fulfilled with probability $1 - \eta$.

At present neither the specific form of the summand $\Omega_1(l/h, -(\ln \eta)/h)$ nor the specific form of the factor $\Omega_2(l/h, -(\ln \eta)/h)$ is important. The main point is that the quantity $\Omega_1(l/h, -(\ln \eta)/h)$ tends to zero as l/h increases, while the quantity $\Omega_2(l/h, -(\ln \eta)/h)$ tends to 1. This fact allowed us to establish a method of minimizing the empirical risk for large samples. For any δ there exists a number T such that as long as $l/h > T$, the inequality

$$P(\alpha) < \nu(\alpha) + \delta$$

is fulfilled with probability $1 - \eta$ simultaneously for the whole set of indicator functions $F(x, \alpha)$, and analogously under the same conditions the inequality

$$I(\alpha) < I_{\text{emp}}(\alpha)(1 + \delta)$$

is fulfilled if the class $F(x, \alpha)$ is a class of arbitrary functions. Therefore a small value of the empirical risk assures (with probability $1 - \eta$) a small value of the expected risk.

However, if the same size is small, the summand $\Omega_1(l/h, -(\ln \eta)/h)$ may differ significantly from zero while the factor $\Omega_2(l/h, -(\ln \eta)/h)$ may differ significantly from 1. In this case a function which yields a small value of the empirical risk may not assure a small value for the expected risk. In order to be able to achieve the guaranteed minimum in the case of small samples it is necessary to take into account not only the value of the empirical risk $\nu(\alpha_{\text{emp}})$ (or $I_{\text{emp}}(\alpha_{\text{emp}})$), but also the value of the summand $\Omega_1(l/h, -(\ln \eta)/h)$ (or of the factor $\Omega_2(l/h, -(\ln \eta)/h)$).

In this chapter we shall consider a method for minimizing risk which, unlike the method of minimizing empirical risk, minimizes the upper bound on the risk, (8.2) (or (8.3)), over both summands (or factors)

$$\nu(\alpha) + \Omega_1\left(\frac{l}{h}, \frac{-\ln \eta}{h}\right)\left(I_{\text{emp}}(\alpha)\Omega_2\left(\frac{l}{h}, \frac{-\ln \eta}{h}\right)\right),$$

—rather than over one summand $\nu(\alpha)$ (or factor $I_{\text{emp}}(\alpha)$). This idea is implemented by the method which we shall call the *method of structural risk minimization.*

Let a *structure* be defined on the set of functions $F(x, \alpha)$, i.e., first a minimal subset of elements S_1 is selected, then a subset S_2 containing S_1, etc., and finally the subset S_q which coincides with the whole set:

$$S_1 \subset S_2 \subset \cdots \subset S_q. \tag{8.4}$$

An ordering (8.4) on the set $F(x, \alpha)$ is given *a priori* (before the occurrence of the sample).

Let the structure be defined in such a manner that the capacity h_i of the subset of functions S_i is less than the capacity h_{i+1} of the subset S_{i+1}, i.e.,

$$h_1 < h_2 < \cdots < h_q.$$

For each subset S_i the bound

$$P(\alpha^i_{\text{emp}}) < \nu(\alpha^i_{\text{emp}}) + \Omega_1\left(\frac{l}{h_i}, \frac{-\ln \eta}{h_i}\right) \tag{8.5}$$

is valid with probability $1 - \eta$ provided the set of indicator functions is ordered; and with probability $1 - \eta$ the bound

$$I(\alpha^i_{\text{emp}}) < I_{\text{emp}}(\alpha^i_{\text{emp}})\Omega_2\left(\frac{l}{h_i}, \frac{-\ln \eta}{h_i}\right) \tag{8.6}$$

is valid provided a set of arbitrary functions is ordered; $F(x, \alpha^i_{\text{emp}})$ is an element which yields the minimum of the empirical risk in S_i.

In (8.5) ((8.6)) the first summand (factor) on the right-hand side decreases as i increases, while the second summand (factor) increases.

The method of structural minimization of the risk amounts to finding a subset S_* in which the function $F(x, \alpha^*_{\text{emp}})$, which minimizes the empirical risk, yields a minimal bound on the expected risk, and choosing this function to be the solution. Observe that since for each element S_i the bound (8.5) ((8.6)) is valid and there are q elements in the structure, the bounds are valid with probability $1 - \eta$ simultaneously for all q functions which minimize the empirical risk (each one in its own S_j). Therefore the solution $F(x, \alpha^*_{\text{emp}})$ obtained using the method of structural risk minimization yields a guaranteed minimal bound with probability $1 - q\eta$ for the risk. In other words the inequality

$$P(\alpha^*_{\text{emp}}) < \nu(\alpha^*_{\text{emp}}) + \Omega_1\left(\frac{l}{h_*}, \frac{-\ln \eta}{h_*}\right) \tag{8.7}$$

is valid (the inequality

$$I(\alpha^*_{\text{emp}}) < I_{\text{emp}}(\alpha^*_{\text{emp}})\Omega_2\left(\frac{l}{h_*}, -\frac{\ln \eta}{h_*}\right) \tag{8.8}$$

is valid) with probability $1 - q\eta$.

When implementing the method of structural minimization it is important that the "guarantee" on the obtained bound on the risk be high (equal to

$1 - \eta$). Therefore, setting $\eta^* = \eta/q$ in (8.7) (and in (8.8)) in place of η, we obtain with probability $1 - \eta$ the inequality

$$P(\alpha^*_{\text{emp}}) < \nu(\alpha^*_{\text{emp}}) + \Omega_1\left(\frac{l}{h_*}, \frac{-\ln \eta + \ln q}{h_*}\right) \tag{8.9}$$

$$\left(I(\alpha^*_{\text{emp}}) < I_{\text{emp}}(\alpha^*_{\text{emp}})\Omega_2\left(\frac{l}{h_*}, \frac{-\ln \eta + \ln q}{h_*}\right)\right). \tag{8.10}$$

For structures consisting of a small number of elements ($q < 20$–100), the increase obtained in the upper bound for $F(x, \alpha^*_{\text{emp}})$ as compared with (8.5) (and (8.6)) will be generally insignificant (since $\log q$ rather than q appears in (8.9) and (8.10)). This means that under the least favorable conditions the guaranteed amount of risk for a solution obtained by the method of structural minimization may be only slightly worse than the guaranteed amount of risk obtained by the method of minimizing the empirical risk, while, as we shall see below, in ordinary situations the gain achieved from using the method of structural risk minimization may be quite substantial.

In what follows it will be convenient to view the method of structural risk minimization as a two-stage minimizing procedure: first a function $F(x, \alpha^i_{\text{emp}})$ which minimizes the empirical risk is selected for each element S_i of the structure (8.4), and then from the q selected functions the one which yields the guaranteed minimum for the value of the risk is chosen. Thus two problems arise in implementing the method of structural risk minimization:

(1) How should the structure on the initial class of functions $F(x, \alpha)$ be defined?
(2) What should the algorithm for choosing the second level be?

The definition of a structure on a set of functions $F(x, \alpha)$ is an informal step in the implementation of the method. The structure should reflect the prior information concerning the problem available to an investigator. Functions which, in the investigator's view, approximate the desired one more probably should be assigned to a class S_i with a lower index. Moreover, the more prior information is available, the smaller the classes with low indices should be.

The definition of an algorithm for choosing the second level reflects the ability to estimate the quality of each one of the decision rules selected at the first level. When constructing algorithms for choosing the second level we shall utilize below the bound on the expected risk (6.48) given by

$$P(\alpha) < \nu(\alpha) + 2\,\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{12}}{l}\left(1 + \sqrt{1 + \frac{\nu(\alpha) l}{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{12}}}\right)$$

if the choice is made from indicator functions (for solving a pattern-recognition problem), and the bound in Theorem 7.6.

$$I(\alpha) < \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau a(p)\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{12}}{l}}} \right]_{\infty}$$

if the choice is made from arbitrary functions (for solving a regression-estimation problem). Utilization of these bounds allows us to obtain the best guaranteed solution for a given structure.†

Another idea for constructing algorithms for the second level is connected with the utilization of a procedure called *moving controls.*

§2 Moving-Control Estimators

We shall estimate the quality of a decision rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk

$$I_{\text{emp}}(\alpha) = \frac{1}{l-1}\sum_{i=1}^{l-1}(y_i - F(x_i, \alpha))^2 \tag{8.11}$$

for a given sequence

$$x_1, y_1; \ldots, x_l, y_l$$

using the following device. Exclude from the sequence the first pair x_1, y_1, and obtain a function which minimizes the empirical risk for the remaining $l-1$ elements of the sequence. Let this function be $F(x; \alpha(\widehat{x_1, y_1}, ; \ldots ; x_l, y_l))$.

Here the symbol $\widehat{x_1, y_1}$ indicates that the pair x_1, y_1 is excluded from the sequence. We shall compute the amount of deviation for the excluded pair x_1, y_1:

$$(y_1 - F(x_1; \alpha(\widehat{x_1, y_1}; \ldots ; x_l, y_l)))^2.$$

Next we shall omit the second pair from the sequence (the first pair is retained) and compute the deviation

$$(y_2 - F(x_2; \alpha(x_1, y_1; \widehat{x_2, y_2}; \ldots ; x_l, y_l)))^2;$$

† Here and below we use the assertion of Theorem 7.6 for $p > 2$. When the condition

$$\sup_{\alpha} \frac{\sqrt[p]{(y - F(x, \alpha))^{2p}}}{(y - F(x, \alpha))^2} \leq \tau$$

is fulfilled for $p > 2$ the best rate of convergence in terms of the order of magnitude is attained. However, all the derivations presented in this chapter may be done using Theorem 7.6. for $1 < p \leq 2$.

In this manner we shall compute deviations for all l pairs. We now form the expression

$$T_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i; \alpha(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)))^2 \quad (8.12)$$

and use it as an estimate for the quality of the function $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk (8.11):

$$I(\alpha_{\text{emp}}(x_1, y_1; \ldots; x_{l-1}, y_{l-1})) = \int (y - F(x, \alpha_{\text{emp}}))^2 P(x, y)\, dx\, dy \sim T_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l).$$

Such an estimation procedure is called *moving control.*

The following theorem is valid:

Theorem 8.1. *A moving-control estimator is unbiased, i.e.,*

$$MI(\alpha_{\text{emp}}(x_1, y_1; \ldots; x_{l-1}, y_{l-1})) = MT_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l).$$

PROOF. The proof consists of verifying the following chain of transformations:

$$\begin{aligned}
&M \int \cdots \int \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)))^2 \\
&\quad \times P(x_1, y_1) \cdots P(x_l, y_l)\, dx_1\, dy_1 \cdots dx_l\, dy_l \\
&= M \int \cdots \int \frac{1}{l} \sum_{i=1}^{l} \left[\int (y_i - F(x_i, \alpha(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)))^2 \right. \\
&\quad \left. \times P(x_i, y_i)\, dx_i\, dy_i \right] \\
&\quad \times P(x_1, y_1) \cdots P(x_{i-1}, y_{i-1}) P(x_{i+1}, y_{i+1}) \cdots P(x_l, y_l) \\
&\quad \times dx_1\, dy_1 \cdots dx_{i-1}\, dy_{i-1}\, dx_{i+1}\, dy_{i+1} \cdots dx_l\, dy_l \\
&= M \frac{1}{l} \sum_{i=1}^{l} I(\alpha(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)) \\
&= MI(\alpha_{\text{emp}}(x_1, y_1; \ldots; x_{l-1}, y_{l-1})).
\end{aligned}$$

The theorem is proved. □

Remark. In proving the theorem the properties of the function $F(x, \alpha)$ were nowhere used. Therefore the moving-control procedure defines unbiased estimators for the quality when estimating indicator functions as well as when estimating an arbitrary functional dependence.

Unbiasedness is, however, an insufficient characterization of an estimator. It is also necessary to know its variance D. If the variance of an estimator T_{mc} is known, one can estimate the average quality of a decision rule which minimizes the empirical risk for samples of size l. Namely, with probability $1 - \eta$ the inequality

$$MI(\alpha_{\text{emp}}(x_1, y_1; \ldots; x_{l-1}, y_{l-1})) \leq T_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l) + \sqrt{\frac{D}{\eta}} \tag{8.13}$$

is valid (here $1 - \eta$ is the confidence level with which the inequality should be valid). The bound (8.13) follows from Theorem 8.1 and Chebyshev's inequality $P\{|MT_{\text{mc}} - T_{\text{mc}}| > \sqrt{D/\eta}\} < \eta$.

However, we cannot compute the variance of a moving-control estimator in a sufficiently general setup. Therefore the applicability of a moving-control procedure for estimating the quality of algorithms minimizing the empirical risk is connected with the assumption that if the sample size exceeds the capacity of a class of functions severalfold, then the variance of the estimator is small (and is of order $1/l$ rather than h/l).†

§3 Moving-Control Estimators in Problems of Regression Estimation

We show that for estimating regression in the class of functions linear in their parameters,

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x),$$

a moving-control estimator admits the following equivalent representation:

$$T_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} \frac{(y_i - f_i^{\mathrm{T}}(\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}Y)^2}{(1 - f_i^{\mathrm{T}}(\Phi^{\mathrm{T}}\Phi)^{-1}f_i)^2}, \tag{8.14}$$

† In the particular case when $F(x, \alpha) = \sum_{i=1}^{n-1} \alpha_i x^i + \alpha_0$, then the vector $x = (x^1, \ldots, x^{n-1})^{\mathrm{T}}$ is multinormally distributed, $y = F(x, \alpha_0) + \xi$, and ξ is a normally distributed noise; this assertion will be proved in Section 4. We are assuming that the same order of magnitude for the variance remains in the general case as well: for a pattern-recognition problem when the class $F(x, \alpha)$ has a finite capacity, and for a regression-estimation problem when the class $F(x, \alpha)$ has a finite capacity and the inequality

$$\sup_{x} \frac{\sqrt{M(y - F(x, \alpha))^4}}{M(y - F(x, \alpha))^2} = \tau < \infty$$

is fulfilled.

where

$$\Phi = \left\| \begin{matrix} \varphi_1(x_1) \cdots \varphi_n(x_1) \\ \cdots\cdots\cdots\cdots \\ \varphi_1(x_l) \cdots \varphi_n(x_l) \end{matrix} \right\|,$$

f_i^{T} is the ith row of the matrix Φ, and Y is an l-dimensional column vector of values of y

$$Y = (y_1, \ldots, y_l)^{\mathrm{T}}.$$

The expression $(\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}Y$ in the numerator of (8.14) is an estimator of the vector of parameters α obtained using the method of least squares for the whole sample. The numberator in (8.14) determines the square of the deviation at the point x_i, and the denominator determines the multiplicative correction which arises when we estimate the parameter α from a sample in which the ith pair x_i, y_i is omitted rather than from the whole sample.

The representation (8.14) is remarkable in that it contains only one matrix inversion (rather than l inversions as in the case for the general procedure described in the preceding section). This fact causes the moving-control procedure to be computationally no more complex than computing residuals using the least-squares procedure.

Below, when constructing algorithms for regression estimation, we search for a solution which yields not only the unconditional minimum (8.11) but also a conditional minimum under the restriction

$$\|\alpha\|^2 = \sum_{i=1}^{n} \alpha_i^2 \leq c.$$

Finding such a conditional minimum is a problem which is equivalent to finding the minimum of the functional†

$$I_{\mathrm{emp}}(\alpha : \gamma) = \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2 + \gamma \|\alpha\|^2, \tag{8.15}$$

where γ is a positive constant depending on c (a Lagrange multiplier).

Estimation of the quality of the solution $\alpha = \alpha_\gamma$ which minimizes the functional (8.15) will also be carried out using the moving-control procedure. We find solutions $\alpha_\gamma(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)$ which minimize the functional (8.15) defined on the $l - 1$ pairs (the pair x_i, y_i is excluded, γ is fixed) and form the quantity

$$T_{\mathrm{mc}}^{\gamma}(x_1, y_1; \ldots; x_l, y_l)$$

$$= \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha_\gamma(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)))^2. \tag{8.16}$$

† Recall that according to Theorem 5.5 estimators of this type (ridge regression estimators) have the minimum variance among all the estimators with the same bias vector.

The quantity T^{γ}_{mc} will be an estimate of the quality of the function $F(x, \alpha_\gamma)$ which minimizes the functional (8.15).

An equivalent representation of (8.16) is obtained using the matrix

$$A_\gamma = (\Phi^{\mathrm{T}}\Phi + \gamma I), \qquad I = \begin{vmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{vmatrix}. \tag{8.17}$$

Namely,

$$T^{\gamma}_{\text{mc}}(x_1, y_1; \ldots, x_l, y_l) = \frac{1}{l}\sum_{i=1}^{l} \frac{(y_i - f_i^{\mathrm{T}} A_\gamma^{-1}\Phi^{\mathrm{T}} Y)^2}{(1 - f_i^{\mathrm{T}} A_\gamma^{-1} f_i)^2}. \tag{8.18}$$

For $\gamma = 0$ (8.18) coincides with (8.14).

We shall derive a representation of a moving-control estimator (8.16) in the form (8.18). Denote

$$(A_\gamma - \|f_i\|^{\mathrm{T}}\|f_i\|)^{-1} = B, \tag{8.19}$$

where $\|f_i\|$ is the matrix with all rows equal to zero except the ith. In the ith row of the matrix the vector f_i^{T} is written.

Then the minimum of (8.15) for the sequence without x_i, y_i is attained for the vector

$$\alpha_\gamma(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l) = B(\Phi^{\mathrm{T}} - \|f_i\|^{\mathrm{T}})Y. \tag{8.20}$$

We express the matrix B in terms of A_γ. To do this we rewrite (8.19) in the form

$$I = BA_\gamma - B\|f_i\|^{\mathrm{T}}\|f_i\|. \tag{8.21}$$

In turn we obtain

$$B = A_\gamma^{-1} + B\|f_i\|^{\mathrm{T}}\|f_i\|A_\gamma^{-1} \tag{8.22}$$

from (8.21). Multiplying the left-hand and right-hand sides of the equality (8.22) by $\|f_i\|^{\mathrm{T}}$, we have

$$B\|f_i\|^{\mathrm{T}} = A_\gamma^{-1}\|f_i\|^{\mathrm{T}} + B\|f_i\|^{\mathrm{T}}\|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}}. \tag{8.23}$$

Equation (8.23) yields

$$B\|f_i\|^{\mathrm{T}} = A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}.$$

Substituting this expression for $B\|f_i\|^{\mathrm{T}}$ into (8.22), we arrive at

$$B = A_\gamma^{-1} + A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}\|f_i\|A_\gamma^{-1}. \tag{8.24}$$

We now compute $\alpha_\gamma = \alpha_\gamma(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)$. In view of (8.24) we have

$$\begin{aligned} \alpha_\gamma &= B(\Phi^{\mathrm{T}} - \|f_i\|^{\mathrm{T}})Y \\ &= A_\gamma^{-1}\Phi^{\mathrm{T}}Y + A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}\|f_i\|A_\gamma^{-1}\Phi^{\mathrm{T}}Y \\ &\quad - A_\gamma^{-1}\|f_i\|^{\mathrm{T}}Y - A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}\|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}}Y. \end{aligned}$$

We compute the square of the deviation utilizing the equality

$$f_i^{\mathrm{T}} A_\gamma^{-1} \|f_i\|^{\mathrm{T}} Y = f_i^{\mathrm{T}} A_\gamma^{-1} f_i y_i.$$

We thus obtain

$$\begin{aligned}(y_i - F(x_i, \alpha_\gamma))^2 &= (y_i - f_i^{\mathrm{T}} A_\gamma^{-1} \Phi^{\mathrm{T}} Y - f_i^{\mathrm{T}} A_\gamma^{-1} \|f_i\|^{\mathrm{T}} (I - \|f_i\| A_\gamma^{-1} \|f_i\|^{\mathrm{T}})^{-1} \|f_i\| A_\gamma^{-1} \Phi^{\mathrm{T}} Y \\ &\quad + f_i^{\mathrm{T}} A_\gamma^{-1} f_i y_i + f_i^{\mathrm{T}} A_\gamma^{-1} \|f_i\|^{\mathrm{T}} (I - \|f_i\| A_\gamma^{-1} \|f_i\|^{\mathrm{T}})^{-1} f_i^{\mathrm{T}} A_\gamma^{-1} f_i y_i)^2 \\ &= \left(\frac{y_i}{1 - f_i^{\mathrm{T}} A_\gamma^{-1} f_i} - \left(1 + \frac{f_i^{\mathrm{T}} A_\gamma^{-1} f_i}{1 - f_i^{\mathrm{T}} A_\gamma^{-1} f_i} \right) f_i^{\mathrm{T}} A_\gamma^{-1} \Phi^{\mathrm{T}} Y \right)^2 = \frac{(y_i - f_i^{\mathrm{T}} A_\gamma^{-1} \Phi^{\mathrm{T}} Y)^2}{(1 - f_i^{\mathrm{T}} A_\gamma^{-1} f_i)^2}.\end{aligned}$$

Finally we arrive at

$$T_{\mathrm{mc}}^{\gamma}(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} \left(\frac{y_i - f_i^{\mathrm{T}} A_\gamma^{-1} \Phi^{\mathrm{T}} Y}{1 - f_i^{\mathrm{T}} A_\gamma^{-1} f_i} \right)^2.$$

§4 Estimating the Expected Risk for Samples of Arbitrary Size

In this section the quality of the linear regression

$$y = \sum_{i=1}^{n} \alpha_i x^i + \alpha_0$$

obtained by the least-squares method is estimated. For this purpose we shall construct a parametric family of statistics by which we shall estimate the expected risk for samples of arbitrary size using the sample

$$x_1, y_1; \ldots; x_l, y_l. \tag{8.25}$$

Under certain conditions we shall show that the introduced estimators are unbiased and shall find the bound of their variance. Let us introduce estimators

$$J_p(x_1, y_1; \ldots; x_l, y_l) = \left(\frac{1 + \dfrac{1}{l+p} + \dfrac{n}{l+p-n-2}}{1 + \dfrac{1}{l-1} + \dfrac{n}{l-n-3}} \right) T_{\mathrm{mc}}(x_1, y_1; \ldots; x_l, y_l),$$

$$p > -l + n + 2, \qquad l > n + 3,$$

where

$$J_{-1}(x_1, y_1; \ldots; x_l, y_l) = T_{\mathrm{mc}}(x_1, y_1; \ldots; x_l, y_l)$$

is an estimator of the moving control.

Using the estimator $J_p(x_1, y_1; \ldots; x_l, y_l)$ we shall determine $MI(\alpha_{\text{emp}}(x_1, y_1; \ldots; x_{l+p}, y_{l+p}))$, the expected risk for samples of size $l + p$. The statistics $J_p(x_1, y_1; \ldots; x_l, y_l)$ allow us to find the important specifics of the problem in hand: for instance, to estimate the minimum possible risk for our problem (the minimum risk is estimated as $J_\infty(x_1, y_1; \ldots; x_l, y_l)$) or to estimate the reduction in risk if r elements are added to the sample (8.25) (the reduction in risk is estimated as $J_0(x_1, \theta_1; \ldots, x_l, y_l) - J_r(x_1, y_1 \ldots; x_l, y_l)$).

The following theorem holds

Theorem 8.2. *Let the probability distribution on pairs x, y be such that y_j is related to x_j as follows:*

$$y_j = \sum_{i=1}^{n} \alpha_i x_j^i + \alpha_0 + \xi_j,$$

where ξ_j is a random variable independent of x distributed according to $N(0, \sigma^2)$ and x is a random n-dimensional vector distributed according to $N(\mu, \Sigma)$.

Then for any $l > n + 7$ and any $p > -l + n + 2$ the statistic $J_p(x_1, y_1; \ldots; x_l, y_l)$ determines an unbiased estimator of the expected risk $MI(\alpha_{\text{emp}}(x_1, y_1; \ldots; x_{l+p}, y_{l+p}))$. The variance† of the estimator is bounded by the inequality

$$\frac{2\sigma^4}{l}\left(\frac{1 + \frac{1}{l+p} + \frac{n}{l+p-n-2}}{1 + \frac{1}{l-1} + \frac{n}{l-n-3}}\right)^2 A < D(J_p(\cdot)) < \frac{2\sigma^4}{l}\left(\frac{1 + \frac{1}{l+p} + \frac{n}{l+p-n-2}}{1 + \frac{1}{l-1} + \frac{n}{l-n-3}}\right)^2 B, \quad (8.26)$$

where

$$A = \left(1 + \frac{1}{l-1} + \frac{n}{l-n-3}\right)^3 - \frac{nl^2}{(l-n-5)^3},$$

$$B = \left(1 + \frac{1}{l-1} + \frac{n+4}{l-n-7}\right)^3 + \frac{nl^2}{(l-n-5)^3}.$$

Corollary 1. *The variance of the moving-control estimator is bounded by the inequalities*

$$\frac{2\sigma^4}{l} A < D(T_{\text{mc}}) < \frac{2\sigma^4}{l} B.$$

† We denote the variance of a random variable z by $D(z)$.

Corollary 2. *The root-mean-square relative error in estimating the expected risk* $MI(\alpha_{\text{emp}}(x_1, y_1; \ldots; x_{l+p}, y_{l+p}))$ *is independent of* p *and is bounded by the inequality*

$$\frac{\sqrt{\frac{2A}{l}}}{1 + \frac{1}{l-1} + \frac{n}{l-n-3}} < \frac{\sqrt{D(J_p(x_1, y_1; \ldots; x_l, y_l))}}{MI(\alpha_{\text{emp}}(x_1, y_1; \ldots; x_{l+p}, y_{l+p}))} < \frac{\sqrt{\frac{2B}{l}}}{1 + \frac{1}{l-1} + \frac{n}{l-n-3}}.$$

Note that unlike Theorem 8.1 where the estimator $T_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l)$ is said to be unbiased for any model for estimating dependences, Theorem 8.2 asserts that the estimators $J_p(x_1, y_1; \ldots; x_l, y_l)$, $p > -l + n + 2$, are unbiased only for the special case of estimating the linear regression by the least-squares method.

To clarify the estimation of the variance (8.26) let us consider two random independent samples, one of size $l + p$:

$$R : x_1, y_1; \ldots; x_{l+p}, y_{l+p},$$

and another of size k:

$$R^* : x_1^*, y_1^*; \ldots; x_k^*, y_k^*.$$

Using the least-squares method, we estimate the parameters α_{emp} from the sample R, and then we estimate the quality of the estimated regression from the sample R^*:

$$I_{\text{emp}}^{R^*}(\alpha_{\text{emp}}(R)) = \frac{1}{k} \sum_{i=l+1}^{l+k} (y_i^* - \alpha_{\text{emp}}^{\mathrm{T}}(R) x_i^*)^2.$$

Clearly the random variable $I_{\text{emp}}^{R^*}(\alpha_{\text{emp}}(R))$ is an unbiased estimator of the quality of the algorithm for regression estimation, i.e.,

$$M_R M_{R^*} I_{\text{emp}}^{R^*}(\alpha_{\text{emp}}(R)) = M_R I(\alpha_{\text{emp}}(R)).$$

(Here M_R (M_{R^*}) indicates mathematical expectation with respect to samples R (R^*).) Therefore the accuracy of this estimator of the quality of the algorithm is determined by the variance of the random variable $I_{\text{emp}}^{R^*} D(\alpha_{\text{emp}}(R))$. Below we shall compute the value of this variance (cf. Equation (8.44).) It turns out that the variance satisfies

$$D(I_{\text{emp}}^{R^*}(\alpha_{\text{emp}}(R))) > \frac{2\sigma^4}{k}\left(1 + \frac{n}{l+p-n-2}\right)^2 + \frac{2\sigma^4 n}{(l+p-n-2)^2}, \quad l + p > n + 2 \quad (8.27)$$

Comparing the bounds (8.26) and (8.27), we conclude that the moving-control estimator has approximately the same precision as the estimator $I^{R^*}_{\text{emp}}(\alpha_{\text{emp}}(R))$ obtained from the sample R of size $l + p$ and the sample R^* of size $k = l - n$.

The proof of the theorem is based on the following two lemmas.

Lemma 8.1. *Let A be a symmetric $n \times n$ matrix, and ξ be an n-dimensional vector distributed according to $N(0, \sigma^2 I)$. Then the equality*

$$M_\xi(\xi^T A\xi)^2 = \sigma^4[2 \operatorname{Sp} A^2 + (\operatorname{Sp} A)^2]$$

is fulfilled.

PROOF. We write the expression $M_\xi(\xi^T A\xi)^2$ termwise assuming that $A = |a_{ij}|$. We obtain

$$\begin{aligned} M_\xi(\xi^T A\xi)^2 &= M_\xi \sum_{i,j,s,t} a_{ij}a_{st}\xi_i\xi_j\xi_s\xi_t \\ &= M_\xi\left[2\sum_{i\neq j} a_{ij}^2\xi_i^2\xi_j^2 + \sum_{i\neq j} a_{ii}a_{jj}\xi_i^2\xi_j^2 + \sum_i a_{ii}^2\xi_i^4\right] \\ &= \sigma^4\left[2\sum_{i=j} a_{ij}^2 + \sum_{i\neq j} a_{ii}a_{jj} + 3\sum_i a_{ii}^2\right] \\ &= \sigma^4\left[2\sum_{i,j} a_{ij}^2 + \sum_{i,j} a_{ii}a_{jj}\right]. \end{aligned}$$

Since the matrix A is symmetric ($a_{ij} = a_{ji}$) we have

$$\begin{aligned} M_\xi(\xi^T A\xi)^2 &= \sigma^4\left[2\sum_i\sum_j a_{ij}a_{ji} + \left(\sum_i a_{ii}\right)^2\right] \\ &= \sigma^4[2 \operatorname{Sp} A^2 + (\operatorname{Sp} A)^2], \end{aligned}$$

q.e.d. □

Lemma 8.2. *Let the random n-dimensional vector $\zeta = (\zeta_1, \ldots, \zeta_n)^T$ and a random $n \times n$ matrix H be statistically independent. Let the vector ζ be normally distributed $N(0, I)$ and the matrix H be distributed according to the Wishart distribution $W_{l,n}(H, I)$.*

Then the random variable

$$\varphi = \zeta^T H^{-1}\zeta - \frac{(\bar{1}^T H^{-1}\zeta)^2}{\bar{1}^T H^{-1}\bar{1}}, \qquad \bar{1} = (1, 0, \ldots 0)^T$$

is distributed as the composition

$$\varphi = \frac{\beta}{u}$$

of two independent random variables β and u, where β is a χ^2 random variable with $n - 1$ degrees of freedom ($\beta \sim \chi^2_{n-1}$) and u is a χ^2 random variable with $l - n + 1$ degrees of freedom ($u \sim \chi^2_{l-n+1}$).

PROOF. We introduce an orthogonal matrix B_T of dimension $n \times n$ with the first row given by $\bar{1}^T = (1, 0, \ldots, 0)$ and the second by the vector

$$\left(0, \frac{\zeta_2}{\sqrt{\sum_{i=1}^{n} (\zeta_i)^2}}, \ldots, \frac{\zeta_n}{\sqrt{\sum_{i=1}^{n} (\zeta_i)^2}}\right);$$

the remaining $n-2$ rows are given by vectors which form an orthonormal system with the first two vectors.

For this matrix the equality

$$B\zeta = (\alpha, \beta, 0, \ldots, 0)^T \tag{8.28}$$

is valid, where α, β are independent random variables, the first being distributed according to $N(0, 1)$ normal distribution while the second is a χ^2 variable with $n-1$ degrees of freedom.

Since a quadratic form remains invariant under an orthogonal transformation, it follows that

$$\alpha = \zeta^T H^{-1} \zeta = (B\zeta)^T (BHB^T)^{-1} (B\zeta),$$

$$b = \frac{(\bar{1}^T H^{-1} \zeta)^2}{\bar{1}^T H^{-1} \bar{1}} = \frac{[(B\bar{1})^T (BHB^T)]^{-1} (B\zeta)]^2}{(B\bar{1})^T (BHB^T)^{-1} (B\bar{1})}.$$

Denote $BHB^T = C$, and decompose the inverse matrix C^{-1} into blocks

$$C^{-1} = \left|\begin{array}{c|c} C_{11}^{-1} & C_{12}^{-1} \\ \hline C_{21}^{-1} & C_{22}^{-1} \end{array}\right|, \qquad C_{11}^{-1} = \begin{vmatrix} c^{11} & c^{12} \\ c^{21} & c^{22} \end{vmatrix}.$$

Taking (8.28) into account, we thus obtain

$$\varphi = a - b = \beta \frac{c^{22} c^{11} - (c^{12})^2}{c^{11}}. \tag{8.29}$$

It is known [5, Theorem 4.33] that the matrix $C_2 = (C_{11}^{-1})^{-1}$ is a 2×2 matrix distributed according to the Wishart distribution $W_{l-n+1,2}(C_2 I)$. This matrix therefore admits the following representation in terms of independent random variables w, v, u [75a]:

$$C_2 = (C_{11}^{-1})^{-1} = \begin{vmatrix} v^2 + w & v\sqrt{u} \\ v\sqrt{u} & u \end{vmatrix},$$

where

$$v \sim N(0, 1), \qquad w \sim \chi^2_{l-n}, \qquad u \sim \chi^2_{l-n+1}.$$

From this we obtain that the matrix C_{11}^{-1} possesses the following representation:

$$C_{11}^{-1} = \begin{vmatrix} \dfrac{1}{w} & \dfrac{v}{w\sqrt{u}} \\ \dfrac{v}{w\sqrt{u}} & \dfrac{v^2 + w}{uw} \end{vmatrix}.$$

Thus the elements c^{11}, $c^{12} = c^{21}$, c^{22} can be represented as a composition of independent random variables

$$c^{11} = \frac{1}{w}, \qquad c^{12} = \frac{v}{w\sqrt{u}}, \qquad c^{22} = \frac{v^2 + w}{wu}.$$

Substituing the values c^{11}, c^{12}, c^{22} into (8.29), we obtain the assertion of the lemma.

□

PROOF OF THE THEOREM. To simplify the notation, we shall assume that the vector x consists of $n+1$ coordinates $x = (q, x^1, \ldots, x^n)$, where $q > 0$ is a fixed number and the vector α consists of numbers $\alpha = (\alpha_0/q, \alpha_1, \ldots, \alpha_n)^T$. In this notation the linear regression is written in the form $y = \alpha^T x$. Note that according to Theorem 8.1

$$MI(\alpha_{\text{emp}}(x_1, y_1; \ldots; x_{l+p}, y_{l+p}) = MT_{\text{mc}}(x_1, y_1; \ldots; x_{l+p+1}, y_{l+p+1}).$$

The estimator $J_p(x_1, y_1; \ldots; x_l, y_l)$ will be proved unbiased if we show that

$$MT_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l) = \sigma^2\left(1 - \frac{1}{l-1} + \frac{n}{l-n-3}\right)$$

In this case the equality

$$\begin{aligned} MT_{\text{mc}}&(x_1, y_1; \ldots; x_{l+p+1}, y_{l+p+1}) \\ &= \left(\frac{1 + \dfrac{1}{l+p} + \dfrac{n}{l+p-n-2}}{1 + \dfrac{1}{l-1} + \dfrac{n}{l-n-3}}\right) MT_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l) \\ &= MJ_p(x_1, y_1; \ldots; x_l, y_l) \end{aligned}$$

holds.

Furthermore, it is obvious that

$$D(J_p(x_1, y_1; \ldots; x_l, y_l)) = \left(\frac{1 + \dfrac{1}{l+p} + \dfrac{n}{l+p-n-2}}{1 + \dfrac{1}{l-1} + \dfrac{n}{l-n-3}}\right)^2 \times D(T_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l)).$$

Consequently to prove Theorem 8.2 it is sufficient to show that the equality

$$MT_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l) = \sigma^2\left(1 + \frac{1}{l-1} + \frac{3}{l-n+3}\right)$$

holds, as does the inequality

$$\frac{2\sigma^4}{l} A < D(T_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l)) < \frac{2\sigma^4}{l} B,$$

where

$$A = \left(1 + \frac{1}{l-1} + \frac{n}{l-n-3}\right)^3 - \frac{nl^2}{(l-n-5)^3},$$

$$B = \left(1 + \frac{1}{l-1} + \frac{n+4}{l-n-7}\right)^3 + \frac{nl^2}{(l-n-5)^3}.$$

In other words, it is sufficient to calculate the mathematical expectation and estimate the variance of the moving control.

(1) We shall calculate the mathematical expectation and bound the variance of the random variable

$$T_{\text{mc}} = \frac{1}{l} \sum_{i=1}^{l} \frac{(y_i - x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y)^2}{(1 - x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i)^2}$$

which forms an estimate of the moving control for a linear regression. Taking into account that $y_i = \alpha^{\mathrm{T}}x_i + \xi_i$, we obtain

$$T_{\text{mc}} = \frac{1}{l} \sum_{i=1}^{l} \frac{(\xi_i - x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\xi)^2}{(1 - x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i)^2}, \tag{8.30}$$

where

$$\xi = (\xi_1, \ldots, \xi_l)^{\mathrm{T}}.$$

Now introduce a diagonal matrix $\mathscr{B}$ with nonzero elements

$$b_{ii} = (1 - x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i)^2,$$

and rewrite (8.30) in the form

$$T_{\text{mc}} = \frac{1}{l} \xi^{\mathrm{T}}(I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})\mathscr{B}^{-1}(I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})\xi,$$

where I is the unit matrix. By definition,

$$MT_{\text{mc}} = M_X M_\xi(T_{\text{mc}} \mid X),$$

$$D(T_{\text{mc}}) = M_X M_\xi(T_{\text{mc}}^2 \mid X) - (M_X M_\xi(T_{\text{mc}} \mid X))^2.$$

We compute the quantities

$$M_\xi(T_{\text{mc}} \mid X), \qquad M_\xi(T_{\text{mc}}^2 \mid X).$$

Elementary calculations yield

$$M_\xi(T_{\text{mc}} \mid X) = \frac{\sigma^2}{l} \sum_{i=1}^{l} \frac{1}{1 - \gamma_i}, \tag{8.31}$$

where

$$\gamma_i = x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i.$$

Now compute the quantity

$$M_\xi(T_{\text{mc}}^2 \mid X) = \frac{1}{l^2} M_\xi(\xi^{\mathrm{T}}(I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})\mathscr{B}^{-1}(I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})\xi)^2.$$

For this purpose we use the result of Lemma 8.1, the facts that $\mathrm{Sp}(AB) = \mathrm{Sp}(BA)$, $\mathrm{Sp}(A^2) \leq \mathrm{Sp}(A^{\mathrm{T}}A)$ and the property

$$(I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})(I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}) = (I - X(X^{\mathrm{T}}X)X^{\mathrm{T}}).$$

We obtain

$$\frac{1}{l^2} M_\xi(\xi^T(I - X(X^TX)^{-1}X^T)\mathscr{B}^{-1}(I - X(X^TX)^{-1}X^T)\xi)^2$$

$$\leq \frac{\sigma^4}{l^2}\{2\,\mathrm{Sp}[\mathscr{B}^{-1}(I - X(X^TX)^{-1}X^T)\mathscr{B}^{-1}]$$

$$+ (\mathrm{Sp}[(I - X(X^TX)^{-1}X^T)\mathscr{B}^{-1}])^2\}$$

$$= \frac{2\sigma^4}{l^2}\sum_{i=1}^{l}\frac{1}{(1-\gamma_i)^3} + \frac{\sigma^4}{l^2}\left(\sum_{i=1}^{l}\frac{1}{1-\gamma_i}\right)^2. \tag{8.32}$$

(2) Let X_i be a matrix obtained from X by deletion of the ith row. The following equality is valid:

$$\frac{1}{1 - x_i^T(X^TX)^{-1}x_i} = 1 + x_i^T(X_i^TX_i)^{-1}x_i.$$

This equality follows from Bartlett's formula for a symmetric matrix K of dimension $(n+1)\times(n+1)$ and $(n+1)$-dimensional vectors f and h:

$$f^T(K + hh^T)^{-1}f = f^TK^{-1}f - \frac{(f^TK^{-1}h)(h^TK^{-1}f)}{1 + h^TK^{-1}h}, \tag{8.33}$$

if we set

$$K = X_i^TX_i, \qquad h = x_i, \qquad f = x_i.$$

Using the notation $\gamma_i^* = x_i^T(X_i^TX_i)^{-1}x_i$ and taking into account that

$$\frac{1}{1-\gamma_i} = 1 + \gamma_i^*$$

we rewrite (8.31) and (8.32) in the form

$$M_\xi(T_{\mathrm{mc}}|X) = \frac{\sigma^2}{l}\sum_{i=1}^{l}(1+\gamma_i^*),$$

$$M_\xi(T_{\mathrm{mc}}^2|X) \leq \frac{2\sigma^4}{l^2}\sum_{i=1}^{l}(1+\gamma_i^*)^3 + \frac{\sigma^4}{l^2}\left(\sum_{l=1}^{l}(1+\gamma_i^*)\right)^2.$$

Next we obtain

$$MT_{\mathrm{mc}} = \sigma^2 M_X(1+\gamma_i^*), \tag{8.34}$$

$$D(T_{\mathrm{mc}}) \leq \frac{\sigma^4}{l}2M_X(1+\gamma_i^*)^3 + D_X(\gamma_i^*)] + \sigma^4\frac{l-1}{l}\rho D_X(\gamma_i^*), \tag{8.35}$$

where

$$\rho = \frac{M_X(\gamma_i^*\gamma_j^*) - M_X(\gamma_i^*)M_X(\gamma_j^*)}{D_X(\gamma_i^*)}$$

is the correlation coefficient between the variables γ_i^* and γ_j^*. Since $-1 \leq \rho \leq 1$, we have

$$D(T_{\mathrm{mc}}) \leq \frac{2\sigma^4}{l}M_X(1+\gamma_i^*)^3 + \sigma^4[M_X(\gamma_i^*)^2 - (M_X(\gamma_i^*))^2],$$

$$D(T_{\mathrm{mc}}) > \frac{2\sigma^4}{l}M_X(1+\gamma_i^*)^3 - \sigma^4[M_X(\gamma_i^*)^2 - (M_X(\gamma_i^*))^2]. \tag{8.36}$$

Thus, to obtain the mathematical expectation and to bound the variance, the quantities $M_X(\gamma_i^*)^p$, $p = 1, 2, 3$, should be computed.

(3) We carry out the following construction. For a fixed i we construct from matrix X a new matrix $\hat{X}$, formed by the vectors

$$\hat{x}_j = (q, x_j^1 - x_{\text{cp}}^1(i), \ldots, x_j^{n-1} - x_{\text{cp}}^{n-1}(i))^{\text{T}}, \qquad j = 1, \ldots, l,$$

where

$$x_{\text{cp}}(i) = \frac{1}{l-1} \sum_{\substack{j=1 \\ j \neq i}}^{l} x_j$$

(the summation is not extended over x_i). Clearly the moving-control estimator obtained from the sample

$$\hat{x}_1, y_1; \ldots; \hat{x}_l, y_l \tag{8.37}$$

coincides with the analogous estimator obtained from the sample (8.25).

Next construct from $\hat{X}$ a new matrix Z which differs from $\hat{X}$ only in the first column. This matrix is formed by the vectors

$$z_j = (q + x_j^0 - x_{\text{cp}}^0(i), x_j^1 - x_{\text{cp}}^1(i), \ldots, x_j^n - x_{\text{cp}}^n(i)), \qquad j = 1, 2, \ldots, l,$$

where x^0 is a normal $N(0, 1)$ random variable independent of x.

For q sufficiently large the moving control estimator computed from the sample

$$z_1, y_1; \ldots; z_l, y_l \tag{8.38}$$

differs from the estimator computed from (8.37) by an amount of the order $1/q$. Therefore, setting $q \to \infty$, we assume that the estimator based on (8.38) coincides with the estimator obtained from (8.25).

Introduce the notation $\zeta_j^{m+1} = x_j^m - x_{\text{cp}}^m(i)$, $m = 0, \ldots, n$,

$$\zeta_j = (\zeta_j^1, \ldots, \zeta_j^n)^{\text{T}}, \qquad h = (q, 0, \ldots, 0)^{\text{T}}.$$

In this notation the vectors z can be represented as

$$z_j = h + \zeta_j,$$

where ζ is an n-dimensional random vector distributed according to $N(0, \Sigma')$. Moreover the covariance matrix Σ' may be arbitrary, but since the random variable γ_i^* is invariant with respect to rotation and change in scale of coordinates, we may assume when estimating γ_i^* that

$$z_j = h + \zeta_j^*, \qquad \sum_{\substack{j=1 \\ j \neq i}}^{l} \zeta_j^* = 0,$$

where the vectors ζ^* are distributed according to the $N(0, I)$ distribution.

Thus the equality

$$z_i = U_i + N_i$$

is valid, where

$$U = \begin{vmatrix} \zeta_1^{*1} & \cdots & \zeta_l^{*n} \\ \cdots & \cdots & \cdots \\ \zeta_l^{*1} & \cdots & \zeta_l^{*n} \end{vmatrix}, \qquad N = \begin{vmatrix} q & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ q & 0 & \cdots & 0 \end{vmatrix};$$

(recall that A_i denotes a matrix obtained from A by omitting the ith row). Furthermore the equality

$$\frac{1}{l-1} Z_i^{\mathrm{T}} Z_i = \frac{1}{l-1} U_i^{\mathrm{T}} U_i + hh^{\mathrm{T}}$$

is valid. We denote

$$\frac{1}{l-1} U_i^{\mathrm{T}} U_i = K_i$$

and utilize Bartlett's formula

$$z_i^{\mathrm{T}}(K_i + hh^{\mathrm{T}})^{-1} z_i = z_i^{\mathrm{T}} K_i^{-1} z_i - \frac{(z_i^{\mathrm{T}} K_i^{-1} h)(h^{\mathrm{T}} K_i^{-1} z_i)}{1 + h^{\mathrm{T}} K_i^{-1} h} \tag{8.39}$$

for

$$(l-1)\gamma_i^* = z_i^{\mathrm{T}}(K_i + hh^{\mathrm{T}})^{-1} z_i.$$

Utilizing the representation of the vector $z_i = h + \zeta_i^*$, we obtain

$$(l-1)\gamma_i^* = \frac{q^2 k^{11} + 2hK_i^{-1}\zeta_i^* + (\zeta_i^*)^{\mathrm{T}} K_i^{-1}\zeta_i^* + q^2 k^{11} K_i^{-1}\zeta_i^* - (h^{\mathrm{T}} K_i^{-1}\zeta_i^*)^2}{1 + q^2 k^{11}}, \tag{8.40}$$

where k^{11} is the first entry of the matrix K_i^{-1}.

Since $q \to \infty$ we have, up to the order of magnitude $1/q^2$,

$$\gamma_i^* = \frac{1}{l-1}\left[(\zeta_i^*)^{\mathrm{T}} K_i^{-1}\zeta_i^* - \frac{(h^{\mathrm{T}} K_i^{-1}\zeta_i^*)^2}{h^{\mathrm{T}} K_i^{-1} h}\right] + \frac{1}{l-1}.$$

Observe that the quantity

$$\frac{1}{l-1}\left[(\zeta_i^*)^{\mathrm{T}} K_i^{-1}\zeta_i^* - \frac{(h^{\mathrm{T}} K_i^{-1}\zeta_i^*)^2}{h^{\mathrm{T}} K_i^{-1} h}\right] \tag{8.41}$$

satisfies the conditions of Lemma 8.2. Indeed, the matrix $(l-1)K_i = H$ is distributed according to the Wishart distribution $W_{l-1,n+1}(H, I)$, and the vector ζ_i^* does not depend on H and is distributed normally $N(0, I)$. Therefore, taking into account Lemma 8.2, we obtain

$$\gamma_i^* = \frac{\beta}{u} + \frac{1}{l-1},$$

where β and u are independent χ^2-distributed random variables with n and $l-n-1$ degrees of freedom respectively.

Utilizing this fact, one can easily deduce from (8.34) and (8.36) that

$$\begin{aligned} M T_{\mathrm{mc}} &= \sigma^2\left(1 + \frac{1}{l-1} + \frac{n}{l-n-3}\right), \\ D(T_{\mathrm{mc}}) &< \frac{2\sigma^4}{l}\left[\left(1 + \frac{1}{l-1} + \frac{n+4}{l-n-7}\right)^3 + \frac{nl^2}{(l-n-5)^3}\right], \\ D(T_{\mathrm{mc}}) &> \frac{2\sigma^4}{l}\left[\left(1 + \frac{1}{l-1} + \frac{n}{l-n-3}\right)^3 - \frac{nl^2}{(l+n-5)^3}\right]. \end{aligned} \tag{8.42}$$

The theorem is proved. □

It remains to compute the variance of $I^{R^*}_{\text{emp}}(\alpha_{\text{emp}})$. Analogously to the proof of the theorem we obtain

$$D(I^{R^*}_{\text{emp}}(\alpha_{\text{emp}}(R))) = \frac{2\sigma^4}{k} M(1+\mu)^2 + \sigma^4 D(\mu), \tag{8.43}$$

where $\mu = x^{*\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x^*$. Here X is the matrix of elements x of the sample by means of which the parameters of the regression are computed, while x^* is an element of the sample by means of which the quality of the estimated function is estimated.

Then, as in the proof of the theorem, we obtain that random variable μ is distributed as the composition

$$\mu = \frac{\beta}{u^*} + \frac{1}{l+p},$$

where β and u^* are independent χ^2 distributed random variables with n and $l+p-n$ degrees of freedom respectively.

Consequently, appropriate computations with (8.43) yield

$$\begin{aligned} D(I^{R^*}_{\text{emp}}(\alpha_{\text{emp}}(R))) = \frac{2\sigma^4}{k}\Bigg[\left(\frac{l+p+1}{l+p}\right)^2 + 2\frac{l+p+1}{l+p}\frac{n}{l+p-n-2} \\ + \frac{n(n+2)}{(l+p-n-2)(l+p-n-4)}\Bigg] \\ + \frac{2\sigma^4 n(l+p-2)}{(l+p-n-2)^2(l+p-n-4)}. \end{aligned} \tag{8.44}$$

Remark. $J_p(x_1, y_1; \ldots; x_l, y_l)$ are not unique unbiased estimators of expected risk for samples of size $l+p$. As in the proof of Theorem 8.2, the estimators

$$\begin{aligned} J^*_p(x_1, y_1; \ldots; x_l, y_l) = \left(1 + \frac{1}{l+p} + \frac{n}{l+p-n-2}\right) \\ \times \frac{1}{l}\sum_{i=1}^{l} \frac{(y_i - x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y)^2}{1 - x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i}, \end{aligned}$$

$$\begin{aligned} J^{**}_p(x_1, y_1; \ldots; x_l, y_l) = \left(1 + \frac{1}{1+p} + \frac{n}{l+p-n-2}\right)\left(1 + \frac{n+1}{l-n-1}\right) \\ \times \frac{1}{l}\sum_{i=1}^{l}(y_i - x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y)^2, \qquad p > -l+n+2, \end{aligned}$$

can be shown to be unbiased. The variances of these estimators are

$$D(J^*_p(\cdot)) = \frac{2\sigma^4}{l}\left(1 + \frac{1}{l+p} + \frac{n}{l+p-n-2}\right)^2\left(1 + \frac{1}{l-1} + \frac{n}{l-n-3}\right),$$

$$D(J^{**}_p(\cdot)) = \frac{2\sigma^4}{l}\left(1 + \frac{1}{l+p} + \frac{n}{l+p-n-2}\right)^2\left(1 + \frac{n+1}{l-n-1}\right).$$

Experiments by computer show, however, that the estimators are more stable to the variation in the conditions of Theorem 8.2 than $J^*_p(x_1, y_1; \ldots; x_l, y_l)$ of $J^{**}_p(x_1, y_1; \cdots; x_l, y_l)$. (This is probably because the estimator $T_{\text{mc}}(x_1, y_1; \ldots; x_l, y_l)$ which leads to $J_p(x_1, y_1; \ldots; x_l, y_l)$ remains unbiased for arbitrary models of estimation of dependences (Theorem 8.1).)

§5 Estimation of Indicator Functions in a Class of Linear Decision Rules

We have determined criteria for choosing the second level for the method of structural risk minimization. These are either minimal guaranteed bounds on the risk or minimal estimators obtained using the moving-control procedure. It now remains to determine the structure on the set of functions $F(x, \alpha)$ in order to define algorithms of structural risk minimization.

In this chapter we consider several methods for defining structures on the set of linear decision rules (for pattern-recognition problems) and on a set of functions linear in a parameter (for regression-estimation problems), and we shall construct corresponding algorithms for estimation of dependences.

First consider a pattern-recognition problem. Let a class of linear decision rules be given:

$$F(x, \alpha) = \theta\left(\sum_{i=1}^{n} \alpha_i \varphi_i(x)\right).$$

We arrange features $\varphi_i(x)$ in the order of decreasing prior probabilities of the "usefulness" for classification and define the following structure of linear decision rules:

$$S_1^1 \subset S_2^1 \subset \cdots \subset S_n^1. \tag{8.45}$$

The class S_1^1 consists of the rules such that only the parameter α_1 may differ from zero. The class S_2^1 consists of the rules which may have two parameters α_1 and α_2 different from zero, and so on. Such an ordering has the following meaning. The first class comprises those rules which use only the first feature for recognition, the second class comprises those rules which utilize the first two features and so on. As it was shown in Chapter 6, the index of capacity of each one of these classes equals i, where i is the number of features used.

For such a structure the method of structural risk minimization amounts to choosing a decision rule $F(x, \alpha^*_{\text{emp}})$ which minimizes the functional

$$R_1(\alpha, i) = 2\frac{i\left(\ln\frac{2l}{i} + 1\right) - \ln\frac{\eta}{12}}{l}\left(1 + \sqrt{1 + \frac{\nu(\alpha) l}{i\left(\ln\frac{2l}{i} + 1\right) - \ln\frac{1}{12}}}\right) + \nu(\alpha) \tag{8.46}$$

with respect to i and $F(x, \alpha) \in S_i^1$. With the confidence level $1 - n\eta$ the probability of an erroneous classification using the decision rule obtained does not exceed the minimum attained in (8.46), i.e.,

$$P\{P(\alpha^*_{\text{emp}}) < R_1(\alpha^*_{\text{emp}}, S^1_*)\} > 1 - n\eta \qquad (n < l).$$

The method of defining a structure on a class of linear decision rules considered above requires prior arrangement of features. This is not always easy to accomplish. We shall therefore define yet another structure which will not require prior arrangement of features. We shall include in the class S_i^2 those decision rules which use for classification no more than i features, i.e., we shall consider the structure

$$S_1^2 \subset \cdots \subset S_n^2. \tag{8.47}$$

This structure is constructed in such a manner that $S_p^1 \subset S_p^2$. Clearly the growth function $m^{S_p^2}(l)$ is bounded in terms of the loss function $m^{S_p^1}(l)$:

$$m^{S_p^2}(l) \leq C_n^p m^{S_p^1}(l) < 1.5 C_n^p \frac{l^p}{p!}. \tag{8.48}$$

Thus the method of structural minimization of the risk on the structure (8.47) results in choosing a function $F(x, \alpha_{\text{emp}}^*)$ which minimizes with respect to i and $F(x, \alpha) \in S_i$ the functional

$$R_2(\alpha, i) = 2\,\frac{i\left(\ln \frac{2l}{i} + 1\right) + \ln C_n^i - \ln \frac{\eta}{12}}{l} \times \left(1 + \sqrt{1 + \frac{\nu(\alpha) l}{i\left(\ln \frac{2l}{i} + 1\right) + \ln C_n^i - \ln \frac{\eta}{12}}}\right) + \nu(\alpha). \tag{8.49}$$

For the obtained solution $F(x, \alpha_{\text{emp}}^*)$ the inequality

$$P\{P(\alpha_{\text{emp}}^*) < R_2(\alpha_{\text{emp}}^*, S_*^2)\} > 1 - n\eta \qquad (n < 1)$$

is valid. One can use the moving-control procedure for both types of structures as an algorithm for choosing the second level.

Thus for solving the problem of pattern recognition in a class of linear decision rules the method of structural minimization recommends that one choose an extremal subspace of features (which may depend, in its composition as well as in the number of features, on whether the system of features is arranged or not) and then construct a decision rule on this space which minimizes the empirical risk.

The choice of the extremal space of features for small samples allows us to increase substantially the probability of correct classification of the observed sample (which did not participate in the sample). The possible gain is exhibited in Table 1, obtained in the course of the solution of a problem in medical differential diagnostics. Here the problem number, the sample size, the initial dimensionality of the binary space of features, the dimensionality of the extremal space of features, and the probabilities of erroneous classification in the initial and extremal spaces are presented. The problem was solved using the algorithms to be described in Addendum 1.

Table 1

Problem no.	Sample size	Initial dimension of space of features	Dimension of extremal space	Probability of error in initial space	Probability of error in extremal space
1	114	84	56	0.21	0.14
2	108	92	47	0.18	0.11
3	131	112	51	0.22	0.10
4	240	134	65	0.13	0.07
5	360	196	82	0.15	0.07

§6 Estimation of Regression in a Class of Polynomials

The problem of determining the number of terms in an expansion in an arranged system of functions is one of the central problems in regression theory. A special case of this problem is the *estimation of polynomial regression.*

The problem is as follows: Let a statistical model which associates a quantity y with the variable x be given by

$$y = R(x) + \xi, \tag{8.50}$$

where $R(x)$ is a polynomial of unknown degree and ξ is an error which does not depend on x, with mean zero and finite variance. Observing the pairs

$$x_1, y_1; \ldots; x_l, y_l,$$

it is required to estimate the polynomial $R^*(x)$ which is "close" to $R(x)$. The closeness is measured in the L_p^2 metric:

$$\rho_L(R(x), R^*(x)) = \left(\int (R(x) - R^*(x))^2 P(x)\, dx \right)^{1/2},$$

where $P(x)$ is the density according to which the values of variable x were chosen.

A traditional method for solving this problem exists: one first determines the degree n of the desired polynomial $R(x)$ and then estimates the regression in the class of functions which are expanded in a system of n orthonormal polynomials of degree $1, 2, \ldots, n$. Thus the main problem is to determine the degree of polynomial regression.

This determination is carried out using standard methods of mathematical statistics. These methods are implemented in the simplest manner for the Gauss–Markov model, i.e., under the condition that the values of x are fixed (cf. Section 2 of Chapter 5). Let these values be $x_1, \ldots, x_l$. In this case it can be assumed without loss of generality that

the function $R(x)$ can be expanded in terms of a system of polynomials $R_i(x)$ orthonormal on $x_1, \ldots, x_l$:

$$\frac{1}{l}\sum_{i=1}^{l} R_p(x_i)R_q(x_i) = \begin{cases} 1 & \text{if } p = q, \\ 0 & \text{if } p \neq q. \end{cases}$$

This system of orthonormal polynomials has the remarkable property that in terms of it the regression $R(x)$ can be represented as

$$R(x) = \sum_{p=1}^{n} \alpha_p^0 R_p(x),$$

where

$$\alpha_p^0 = M\frac{1}{l}\sum_{i=1}^{l} R_p(x_i)y_i.$$

The estimators $\hat{\alpha}_p$ of the parameters computed using the least-squares method can be shown to be equal to

$$\hat{\alpha}_p = \frac{1}{l}\sum_{i=1}^{l} R_p(x_i)y_i. \tag{8.51}$$

Thus the problem of determining the degree of the regression consists of accepting (or rejecting) the hypothesis $\alpha_i^0 = 0$ $(i = 1, 2, \ldots, n)$ on the basis of information about the values of $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$.

Note that if the noise ξ in (8.50) is distributed normally $N(0, \sigma^2)$ with mean zero and variance σ^2, then the random variable $\hat{\alpha}_p$ is also distributed normally but with mean α_p^0 and variance $\sigma_1^2 = \sigma^2/l$. In this case for $\alpha_p^0 = 0$ the quantity

$$(\hat{\alpha}_p)^2 = \left(\frac{1}{l}\sum_{i=1}^{l} R_p(x_i)y_i\right)^2 \tag{8.52}$$

is distributed according to the $\sigma_1^2\chi^2$-distribution with one degree of freedom. If the variance of the noise were known, one could use the distribution

$$\left(\frac{\hat{\alpha}_p}{\sigma_1}\right)^2 \sim \chi_1^2$$

to test the hypothesis $M\hat{\alpha}_p = 0$. In this case if the quantity $(\hat{\alpha}_p/\sigma_1)^2$ exceeds $\varkappa(\eta)$ (the value of $\varkappa(\eta)$ is determined from the condition $P\{\chi_1^2 > \varkappa(\eta)\} = \eta$), then the hypothesis $M\hat{\alpha}_p = \alpha_p^0 = 0$ is rejected at a given significance level η; otherwise the hypothesis is accepted.

However, in practice the variance σ^2 of ξ is unknown. Therefore along with (8.52) the statistic

$$\pi^2 = \frac{1}{l}\sum_{i=1}^{l} y_i^2 - \sum_{i=1}^{p} (\hat{\alpha}_i)^2 \tag{8.53}$$

is considered. If one starts with $p = r + 1$ coefficients $\alpha_i^0 = 0$ $(i = r + 1, \ldots, l)$, the statistic (8.53) is then distributed according to the $\sigma_1^2\chi^2$-distribution with $v = l - r - 1$ degrees of freedom.

We form the statistic $\zeta = v\hat{\alpha}_p^2/\pi^2$. This statistic is distributed according to Fisher's F_{1v}-distribution:

$$\zeta = \frac{v\chi_1^2}{\chi_v^2} \sim F_{1,v}. \tag{8.54}$$

This distribution is tabulated in all practical texts in statistics. Thus utilizing the statistic $v\hat{\alpha}_p^2/\pi^2$, one can determine for a given significance level η whether the hypothesis $\alpha_i^0 = 0$ is acceptable; for this purpose it is sufficient to check whether the inequality

$$\zeta > \varkappa(\eta)$$

is fulfilled.

When solving practical problems it is not necessary to construct a system of polynomials orthonormal on $x_1, \ldots, x_l$. It is easy to verify that the statistic

$$\zeta = \frac{R_r - R_{r+1}}{R_{r+1}}(l - r - 1),$$

where R_r is the residual (the value of the minimum for the empirical risk) computed for polynomials of degree r, is also distributed according to the $F_{1,l-r-1}$-distribution. Thus in the case of a normally distributed random error ξ, utilizing the residuals $R_1, \ldots, R_{l-1}$ computed for polynomials of degrees $1, r, \ldots, l-1$ respectively, it is possible to determine the degree of a polynomial regression by means of Fisher's F-criterion (8.54).

However, the classical scheme of estimating polynomial regression, which involves the determination of the true degree of regression and an approximation to regression in a class of polynomials of this degree, can be successfully implemented only when large samples are used. Only for large samples can one assert that the best approximation can be attained for functions which minimize the empirical risk in the class of polynomials whose degree is equal to the true degree of the regression. For small samples the problem of the most appropriate degree of an approximation remains open.†

Below we shall apply the method of structural minimization to solve this problem, but before proceeding to construct the corresponding algorithms we want to emphasize that actually the problem will be solved in a more general setup than the classical one. We shall not assume that regression is a polynomial—it may be a square-integrable function, but the approximating function will be polynomial. Under these conditions it is required to determine an appropriate approximation.

We shall thus solve the problem using the method of structural risk minimization. For this purpose we shall define a structure on the set of polynomials. Observe that the statement of the problem already contains an indication of the special features in defining the structure

$$S_1 \subset \cdots \subset S_n. \tag{8.55}$$

The set S_p consists of polynomials whose degree does not exceed p. Such an ordering of polynomials is "natural" (but not unique). It corresponds to an ordering according to the number of terms in the expansion of a series consisting of elements

$$1, x, x^2, \ldots, x^n, \ldots, \tag{8.56}$$

† For small samples the classical scheme may yield paradoxical results: the more powerful the criterion used for establishing the degree of regression, the worse the final result may be.

arranged in increasing order of degree n. However, another ordering of the terms of the series is possible—for example,

$$x^5, 1, x^4, x^2, x^7, \ldots. \tag{8.57}$$

Ordering of polynomials in accordance with the expansion in terms of the first p terms of the series (8.57) will result in another structure on the set of polynomial dependences.

Thus consider the structure (8.55) defined by means of an expansion in terms of the first members of the series arranged in accordance with (8.56). Moreover, let the restriction

$$\sup_{\alpha} \frac{\sqrt[p]{M(y - F(x, \alpha))^{2p}}}{M(y - F(x, \alpha))^2} \leq \tau \qquad (p > 2)$$

be fulfilled.† Then in view of Theorem 7.6 the inequality

$$I(\alpha) < \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau a(p) \sqrt{\dfrac{r\left(\ln \dfrac{2l}{r} + 1\right) - \ln \dfrac{\eta}{12}}{l}}} \right]_{\infty} \tag{8.58}$$

is fulfilled with probability $1 - \eta$ simultaneously for all polynomials of degree $r - 1$ (all polynomials $F(x, \alpha)$ belonging to S_r). The inequality (8.58) is fulfilled also for the polynomial $F(x, \alpha'_{\text{emp}})$ which minimizes the empirical risk on S_r.

As an approximation to regression we shall choose the function which minimizes the empirical risk on an element S_* of the structure for which the minimum on the right-hand side of the bound (8.58) is attained. Let the minimum be attained for function $F(x, \alpha^*_{\text{emp}})$ and be equal to $R(\alpha^*_{\text{emp}}, S_*)$. Then the assertion

$$P\{I(\alpha^*_{\text{emp}}) < R(\alpha^*_{\text{emp}}, S_*)\} > 1 - n\eta$$

is valid.

The estimation of polynomial regression is highly efficient in practice for a small sample using the method of structural risk minimization.

The result of estimating regression defined by a polynomial of the fifth degree on the interval $[-2, 2]$ is presented in Figure 6. The estimation was carried out based on measurements of a function at 20 randomly chosen points of the interval $[-2, 2]$. The measurements were subject to an error distributed uniformly on the interval $[-a, a]$, where a is the maximal value of the regression on the interval $[-2, 2]$. In the figure both the empirical data (crosses) and the regression (bold line) are shown. The best approximation in the class of polynomials of the 5th degree is given by open circles on

† As was mentioned above, the knowledge of the bound τ is a weaker requirement than the knowledge of the type of error density, which is necessary for estimating the regression polynomial using classical methods.

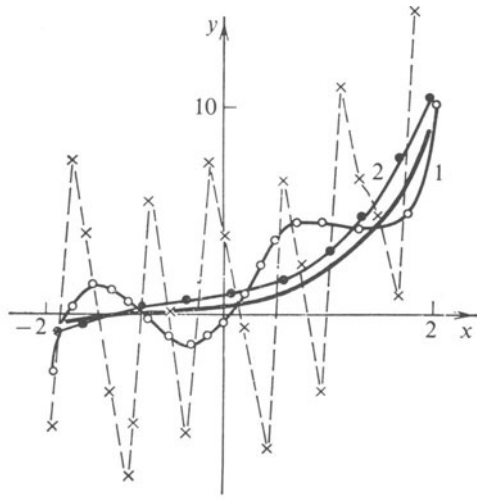

Figure 6

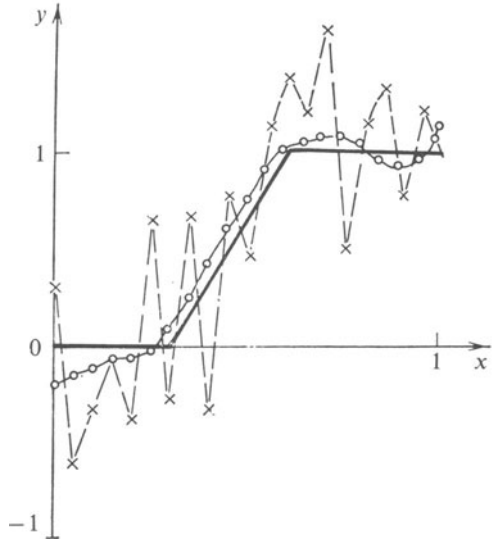

Figure 7

curve 1; the approximation obtained by the method of structural minimization of the risk is a polynomial of the 4th degree—the black dots on curve 2. It can be seen that the approximation of the regression by means of curve 2 is superior to the one given by curve 1. An example of estimating nonpolynomial regression (bold line) in a class of polynomials (thin line) based on 20 observations (crosses) is presented in Figure 7. The functions were estimated using Algorithm D-II.1 described in Addendum II.

§7 Estimation of Regression in a Class of Functions Linear in Their Parameters: Moving Control Method

Consider the class of functions

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x) \tag{8.59}$$

linear in their parameters. There are two approaches to defining a structure on this class:

(1) ordering of functions according to the number of terms in an expansion.
(2) ordering of functions according to the norm of the vector of parameters α (the norm of functions in L_p^2 for a system $\varphi_1(x), \ldots, \varphi_n(x)$ orthonormal in the measure $P(x)$).

We shall construct on these structures algorithms for a structural minimization of the risk which utilize the method of moving control as a criterion for choosing the second level.

(1) *Ordering according to the number of terms in an expansion.* Let an *a priori* arranged system of functions

$$\varphi_1(x), \ldots, \varphi_n(x) \tag{8.60}$$

be given. We shall define on the set of functions $F(x, \alpha)$ the structure

$$S_1 \subset \cdots \subset S_n, \tag{8.61}$$

where the element S_i of the structure contains only those functions which can be expanded in terms of the first i members of the series (8.60). In this case the method of structural minimization involves determination of a subspace $\varphi_1(x), \ldots, \varphi_r(x), 0, \ldots, 0$ of the initial space $\varphi_1(x), \ldots, \varphi_n(x)$ on which the minimum of the quantity

$$T_{\text{mc}}^2(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} \frac{(y_i - (f_i^r)^{\mathrm{T}}(\Phi_r^{\mathrm{T}}\Phi_r)^{-1}\Phi_r^{\mathrm{T}}Y)^2}{(1 - (f_i^r)^{\mathrm{T}}(\Phi_r^{\mathrm{T}}\Phi_r)^{-1}f_i^r)^2} \tag{8.62}$$

is attained. The function $F(x, \alpha_{\text{emp}}^*)$ which minimizes the empirical risk in S_r (the vector of parameters $\alpha_{\text{emp}}^* = (\Phi_r^{\mathrm{T}}\Phi_r)^{-1}\Phi_r^{\mathrm{T}}Y$) is considered to be the best approximation to the regression. In (8.62), (f_i^r) denotes the vector $(\varphi_1(x_i), \ldots, \varphi_r(x_i), 0, \ldots, 0)^{\mathrm{T}}$, and Φ_r is a matrix whose rows are

$$(f_i^r)^{\mathrm{T}} = (\varphi_1(x_i), \ldots, \varphi_r(x_i), 0, \ldots, 0).$$

(2) *Ordering according to the value of the norm of the vector of parameters.* Consider a system of ordered sets

$$S_1 \subset \cdots \subset S_q \tag{8.63}$$

such that subsets S_i contain only functions $F(x, \alpha)$ for which the conditions

$$\sum_{j=1}^{n} \alpha_j^2 \leq c_i \tag{8.64}$$

are fulfilled. The quantities c_i form an increasing sequence:

$$0 < c_1 < c_2 < \cdots < c_q < \infty.$$

One can match to c_i with a monotonically decreasing series of positive quantities γ_i (Lagrange multipliers)

$$\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_q = 0,$$

such that the problem of empirical risk minimization on the set S_r becomes equivalent to the minimization of the functional

$$I_{\gamma_r}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{j=1}^{n} \alpha_j \varphi_j(x_i) \right)^2 + \gamma_r \sum_{j=1}^{n} \alpha_j^2. \tag{8.65}$$

In this case the two-level set-up of the method of structural risk minimization involves choosing at the first level q functions $F(x, \alpha_{\text{emp}}(\gamma_r))$ which minimize for various γ_r the functional (8.65) and then selecting at the second level from the q chosen functions one which yields the minimum for the "moving control" estimator. In other words under this procedure of defining a structure the method of structural minimization first determines γ_r for which the minimum of the expression

$$T_{\text{mc}}^{\gamma_r}(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} \frac{(y_i - f_i^{\mathrm{T}} A_{\gamma_r}^{-1} \Phi^{\mathrm{T}} Y)^2}{(1 - f_i^{\mathrm{T}} A_{\gamma_r}^{-1} f_i)^2} \tag{8.66}$$

is attained (here $A_{\gamma_r} = \Phi^{\mathrm{T}}\Phi + \gamma_r I$) and then determines an $F(x, \alpha^*)$ which minimizes for this γ_r the functional (8.65). This function is defined by the vector of parameters $\alpha^* = A_{\gamma_r}^{-1} \Phi^{\mathrm{T}} Y$.

Finally we shall consider a *combined structure* on the set of linear in parameters functions $F(x, \alpha)$. First we shall order the functions according to the number of terms in the expansion (8.60), and then order each subset S_p consisting of functions expanded into p terms, according to the values of the norm of the vector of parameters (8.64).

Thus we consider the following system of sets

$$\begin{array}{ccccc}
S_{10} & \subset & S_{20} & \subset \cdots \subset & S_{q0} \\
\cup & & \cup & & \cup \\
S_{11} & \subset & S_{21} & \subset \cdots \subset & S_{q1} \\
\cup & & \cup & & \cup \\
\vdots & & \vdots & & \vdots \\
\cup & & \cup & & \cup \\
S_{1n_1} & \subset & S_{2n_2} & \subset \cdots \subset & S_{q,n_q}.
\end{array} \tag{8.67}$$

The element S_{pr} is a subset consisting of functions expanded into p terms of the series and such that the inequality

$$\sum_{i=1}^{p} \alpha_i^2 \leq c_r$$

is fulfilled. The method of structural minimization determines the pair p, γ_r for which the bound on the quality of the algorithm minimizing the empirical functional

$$I_{\gamma_r}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{j=1}^{p} \alpha_j \varphi_j(x_i) \right)^2 + \gamma_r \sum_{j=1}^{p} a_j^2,$$

obtained using the moving-control procedure will be the smallest. Computationally this means that it is required to find a pair p, γ_r for which the minimum of the expression

$$T_{\text{mc}}^{\gamma_r, p}(x_1, y_1; \ldots, x_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} \frac{(y_i - (f_i^p)^{\mathrm{T}} A_{\gamma_r, p}^{-1} \Phi_p^{\mathrm{T}} Y)^2}{(1 - (f_i^p)^{\mathrm{T}} A_{\gamma_r, p}^{-1} f_i^p)^2},$$

$$A_{\gamma_r, p} = (\Phi_p^{\mathrm{T}} \Phi_p + \gamma_r I) \tag{8.68}$$

is attained and to determine the function $F(x, \alpha^*)$ (the vector of parameters $\alpha^* = A_{\gamma_r, p}^{-1} \Phi_p^T Y$).

We have thus described algorithms of structural risk minimization which use the the moving-control procedure as a criterion for choosing the second level. Implementation of these algorithms of regression estimation in a class of functions linear in their parameters turns out to be only slightly more involved than the implementation of the least-squares method. In practice when estimating regression these algorithms yield good and stable results if the sample size is several (2–3) times larger than the dimensionality of the space. The construction of algorithms for structural risk minimization for sample sizes commensurable with (or smaller than) the dimensionality of the parameter vector α is connected with bounds on the probability of a uniform relative deviation of the means from their mathematical expectations.

§8 Estimation of Regression in a Class of Functions Linear in Their Parameters: Uniform Estimating Method

As in the preceding section, we shall consider three types of structures on a class of functions linear in their parameters,

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x):$$

(a) a structure formed according to the number of terms in the expansion;
(b) a structure formed according to the size of the norm of the parameter vector α (the norm of $F(x, \alpha)$ in the L_p^2 metric for an orthogonal—with respect to a measure $P(x)$—system of functions $\varphi_1(x), \ldots, \varphi_n(x)$);
(c) a combined structure formed according to the number of terms in the expansion as well as the size of the norm of the function $F(x, \alpha)$.

Below we shall construct for these structures a method of structural risk minimization based on bounds on the probability of a uniform relative deviation of the means from their mathematical expectations.

(a): For this structure, according to Theorem 7.6, for $p > 2$, the inequality

$$I(\alpha) < \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau a(p) \sqrt{\frac{r\left(\ln \frac{2l}{r} + 1\right) - \ln \frac{\eta}{12}}{l}}} \right]_{\infty} \tag{8.69}$$

is fulfilled with probability $1 - \eta$ simultaneously for all functions in the element S_r of the structure (the set S_r contains functions $F(x, \alpha)$ expanded in terms of the first r members). Since the inequality is valid with probability $1 - \eta$ simultaneously for all functions in S_r, it is fulfilled in particular with probability $1 - \eta$ for the function $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk on S_r. We now choose an element S_* of the structure and the corresponding function minimizing the empirical risk such that the minimum of the bound (8.69) is attained. The function $F(x, \alpha^*_{\text{emp}})$ obtained defines for structure (a) the minimal guaranteed (with probability $1 - \eta n$, where $n < l$ is the number of elements in the structure) value of the risk.

(b): This structure consists of

$$S_1 \subset \cdots \subset S_n. \tag{8.70}$$

Here S_r is the set of functions

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x),$$

for which the relation

$$\sum_{i=1}^{n} \alpha_i^2 \leq c_r$$

is fulfilled. Select on the set S_r a finite ε-net $S_\varepsilon = \{F(x, \alpha_1), \ldots, F(x, \alpha_{N(\varepsilon)})\}$ consisting of $N(\varepsilon)$ elements. According to Theorem 7.5 we have with probability $1 - \eta$ the following bound associated with the function $F(x, \alpha_{\text{emp}})$ which minimizes the value of the empirical risk,

$$I(\alpha_{\text{emp}}) < \left(\varepsilon + \sqrt{\varepsilon^2 + \left[\frac{I_{\text{emp}}(\alpha_i(\alpha_{\text{emp}}))}{1 - T(\varepsilon)} \right]_{\infty}} \right)^2 \tag{8.71}$$

on the sample where

$$T(\varepsilon) = 2\,\tau a(p)\sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln(\eta/24)}{l}}, \qquad p > 2.$$

In the bound (8.71), $F(x, \alpha_i(\alpha_{\text{emp}}))$ is an element of the ε-net closest to $F(x, \alpha_{\text{emp}})$. Thus for a function minimizing the empirical risk on the element S_r of the structure (8.70), a guaranteed bound on the value of the expected risk may be computed. We choose a function (i.e., an element of the structure) for which this bound is minimal.

(c): Each element $S_{q,r}$ of this structure is determined by the number of terms in the expansion

$$F(x, \alpha) = \sum_{i=1}^{q} a_i \varphi_i(x),$$

as well as by the norm of functions

$$\sum_{i=1}^{q} \alpha_i^2 \leq c_r.$$

We set up the method of structural risk minimization for this structure. As a bound on the quality of a function minimizing the empirical risk in $S_{q,r}$, the same bound as above (8.71) is used. We thus choose an element $S_{q,r}$ of the structure and the corresponding function for which the bound is minimal.

In order to construct algorithms of structural risk minimization for structures (b) and (c), it is required to be able to compute the capacity of an ε-net.

§9 Selection of Sample

In this section we shall discuss the concept of *selection of a sample*, which amounts to an exclusion of several elements from the given sample in order to determine, using the remaining set, a function which will yield the smallest guaranteed value for the expected risk.

Note that for problems of pattern recognition the selection of training sequences does not make sense: solutions obtained using minimization of the empirical risk over the whole sample, and over a subsample of it obtained by excluding a minimal number of elements in order that the subsample could be subdivided errorlessly, are obtained for the very same decision rule. This is a corollary of the fact that the loss function $(\omega - F(x, \alpha))^2$ takes only two values, 0 and 1, in pattern-recognition problems. In regression problems, however, the loss function takes on arbitrary positive values, and therefore an exclusion of some elements x and y may substantially change the solution as well as an estimate of the quality of the solution obtained.

Thus let a sample

$$x_1, y_1; \ldots; x_l, y_l \tag{8.72}$$

be given. Consider simultaneously

$$H_l^t = \sum_{m=0}^{t} C_l^m$$

different problems of estimating the functional dependence based on empirical data

$$x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; \widehat{x_j, y_j}; \ldots; x_l, y_l.$$

The notation $\widehat{x_i, y_i}$ indicates that the element (x_i, y_i) has been excluded from (8.72). The problems differ from each other only in that for each of them the functional dependence is estimated through its own sample obtained from (8.72) by excluding at most t elements. (One can construct from (8.72) C_l^m different subsamples consisting of $l - m$ each. Thus there are in all

$$H_l^t = \sum_{m=0}^{t} C_l^m$$

different problems.)

According to Theorem 7.6, for each one of the H_l^t problems the inequality

$$I(\alpha) < \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau a(p) \sqrt{\dfrac{\ln m^S(2(l - t_i)) - \ln(\eta/8)}{l - t_i}}} \right]_\infty$$

is fulfilled with probability $1 - \eta$ (here $t_i \leq t$ is the number of vectors excluded from the ith problem). Consequently, the inequalities

$$I(\alpha) < \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau a(p) \sqrt{\dfrac{\ln m^S(2(l - t_i)) + \ln H_l^t - \ln(\eta/8)}{l - t_i}}} \right]_\infty \tag{8.73}$$

are valid with probability $1 - \eta$ simultaneously for all H_l^t problems. We shall now search for the minimum of the right-hand side of (8.73) over all the H_l^t problems and not only over the elements S_r of the structure and function $F(x, \alpha) \subset S_r$. In other words we shall minimize the functional

$$I_{\text{emp}}(\alpha, \widehat{x_{r_1}, y_{r_1}}; \ldots; \widehat{x_{r_{t_i}}, y_{r_{t_i}}})$$

$$= \left[\frac{\dfrac{1}{l - t_i} \sum_{j=1}^{l}{}^{(t_i)} (y_j - F(x_j, \alpha))^2}{1 - 2\tau a(p) \sqrt{\dfrac{\ln m^{S_r}(2(l - t_i)) + \ln H_l^t - \ln(\eta/8)}{l - t_i}}} \right]_\infty \tag{8.74}$$

over the elements $F(x, \alpha) \in S_r$ and $x_{r_1}, y_{r_1}; \ldots; x_{r_{t_i}}, y_{r_{t_i}}$; here the sign $\sum_{j=1}^{l}{}^{(t_i)}$ indicates that the summation is *not* extended over $t_i \leq t$ elements.

Enumeration over t (usually $t = 1, 2, 3, 4$) yields the smallest value for (8.74). This value determines the guaranteed value (with probability $1 - \eta t$) of the expected risk.

Thus, in searching for the best guaranteed solution—in addition to optimizing over a structure and over functions belonging to elements of the structure—additional optimization over a selection of a subset from a given sample (8.72) can be made. In practice, when the sample is small, the proper selection of this subset from the given set is very often quite useful.

§10 Remarks on a General Theory of Risk Minimization

In this chapter a new principle for minimizing risk in the case of small samples has been formulated. It turns out that if one defines a structure on an admissible set of solutions, then it becomes possible to carry out additional optimization over the elements of the structure. It is only necessary that the structure be given *a priori*. An additional possibility of minimizing risk over the empitical data occurs on account of selecting a sample.

In this chapter we have applied the method of structural minimization of the risk for solving problems of pattern recognition and regression estimation, and the idea of selecting a sample was used in the latter case (in the former case it does not reduce the guaranteed bound on the risk, since the loss function is too simple).

The question arises: how general are the methods of structural risk minimization and of sample selection?

Clearly the method of structural risk minimization is applicable for solving an arbitrary risk-minimization problem for which a bound on a uniform deviation or a relative uniform deviation can be derived (cf. Chapter 6, Section 11, Chapter 7, Section 8). In this case for minimization of the functional

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz$$

on the basis of empirical data $z_1, \ldots, z_l$, a structure

$$\Lambda_1 \subset \Lambda_2 \subset \cdots \subset \Lambda_q$$

is defined on the set Λ of functions $Q(z, \alpha)$.

On each element Λ_i of this structure a value of the parameter $\alpha^i_{\text{emp}} \in \Lambda_i$ which minimizes the empirical risk

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha)$$

is obtained, and then, using the bounds presented in Section 11 of Chapter 6 and Section 8 of Chapter 7, a parameter α^*_{emp} is selected from the q parameters obtained which yields a guaranteed minimum for the value of the expected risk.

A selection of a sample may also be carried out when there exists a uniform bound on the expected risk.

Chapter 9

Solution of Ill-posed Problems. Interpretation of Measurements Using the Method of Structural Risk Minimization

§1 Ill-posed Problems of Interpreting Results of Indirect Experiments

Let it be required to estimate the functional dependence $f(t, \alpha_0) = f(t)$ in the class of functions $f(t, \alpha)$. (Here $f(t)$ belongs to $f(t, \alpha)$.) Moreover let the situation be such that it is impossible to measure directly the values of the function $f(t)$, but one can measure values of another function $F(x)$ $(a \leq x \leq b)$ related to the desired one by means of the operator equation

$$Af(t) = F(x). \tag{9.1}$$

The operator A maps in a one-to-one manner elements $f(t, \alpha)$ of the space E_1 into elements $F(x, \alpha)$ of the space E_2.

Let the following measurements of the function $F(x)$ be taken:

$$x_1, y_1; \ldots; x_l, y_l. \tag{9.2}$$

The pair x_i, y_i denotes that the measured value of the function $F(x_i)$ at point x_i is y_i.

It is required, knowing the operator A and measurements (9.2) to estimate the function $f(t) = f(x, \alpha_0)$ in the class $f(x, \alpha)$. Here it is assumed that the problem of solving the operator equation (9.1) may be ill posed.

We shall estimate the function $f(t)$ in the case when:

(1) the values of function $F(x)$ are measured with an additive error

$$y_i = F(x_i) + \xi, \qquad M\xi = 0, \quad M\xi^2 = \sigma^2 < \infty,$$

which does not depend on x;

(2) the points x_i at which the measurements are taken are chosen randomly and independently according to some nonvanishing density on $[a, b]$. Below we shall assume that this density is uniform.

It was shown in Chapter 1 that the function $f(t, \alpha_0)$ which is the preimage in E_1 of the regression $F(x, \alpha_0)$ in the space E_2—i.e., is the preimage of the point of minimum of the functional

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(y|x)\, dy\, dx \tag{9.3}$$

—coincides with the solution to Equation (9.1). However, it is an impossible task to obtain the (exact) regression from a finite sample. One can only hope to obtain a function $F(x, \hat{\alpha})$ which is close (in a metric of the space E_2) to the regression, and then to choose as a solution of Equation (9.1) the preimage $f(t, \hat{\alpha})$ of this function in the space E_1. Such an approach is not always successful: it is inconsistent in the sense that if Equation (9.1) defines an ill-posed problem, widely different preimages in E_1 may (though not necessarily) correspond to close images in E_2.

In our case it implies that not all methods of risk minimization in the space of images may be utilized for solving the problem of interpreting results of indirect experiments, and that there may exist methods of risk minimization which produce only those elements $F(x, \hat{\alpha})$ in the space E_2 which are images of functions that are close to the desired solution. These methods of risk minimization (if they exist) should be utilized for solving ill-posed problems of interpreting measurements.

Below we shall show that under certain conditions algorithms of structural risk minimization may be utilized for solving ill-posed measurement problems. We shall prove that as the number of measurements increases a sequence of solutions obtained using the method of structural risk minimization converges to the desired function $f(t)$.

§2 Definitions of Convergence

Let a measure of closeness between functions $\rho_{E_1}(f(t, \alpha_1), f(t, \alpha_2)) = \rho_{E_1}(\alpha_1, \alpha_2)$ be chosen in E_1, and let an algorithm of estimating dependence $f(t) = f(t, \alpha_0)$ based on indirect experiments

$$x_1, y_1; \ldots; x_l, y_l \tag{9.4}$$

be fixed. Then for each specific realization (9.4) the function $f(t, \hat{\alpha}_l)$ (i.e., the vector of parameters $\hat{\alpha}_l = \alpha(x_1, y_1; \ldots; x_l, y_l)$) may be obtained, and in this manner the sequence

$$\hat{\alpha}_1, \ldots, \hat{\alpha}_l, \ldots \tag{9.5}$$

is generated. This sequence determines a sequence of numbers

$$\rho_{E_1}(\hat{\alpha}_1, \alpha_0), \ldots, \rho_{E_1}(\hat{\alpha}_l, \alpha_0), \ldots, \tag{9.6}$$

which defines the distance between the parameters $\hat{\alpha}_i$ and α_0. Both (9.5) and (9.6) are random sequences generated by an algorithm A for estimating the dependence $f(t)$ and by a particular outcome (9.4) of the indirect experiment. The investigation of algorithms for regression estimation is thus reduced to a study of the convergence of the sequence (9.6).

There exist different versions of the notion of convergence of random sequences. In this chapter we shall utilize two of them: convergence in probability and convergence with probability 1 (almost surely).

Definition 1. A sequence of random variables $\xi_1, \ldots, \xi_l \ldots$ converges to the variable ξ_0 in probability if for any $\varepsilon > 0$ the probability that the inequality

$$|\xi_l - \xi_0| < \varepsilon$$

will be valid approaches 1 as $l \to \infty$, i.e.,

$$\lim_{l \to \infty} P\{|\xi_l - \xi_0| < \varepsilon\} = 1.$$

We shall denote convergence in probability by $\xi \xrightarrow{P} \xi_0$.

Definition 2. A sequence of random variables $\xi_1, \ldots, \xi_l, \ldots$ converges to the variable ξ_0 with probability 1 if for any $\varepsilon > 0$ the probability that the inequality

$$\sup_{i \geq l} |\xi_i - \xi_0| < \varepsilon$$

will be valid approaches 1 as $l \to \infty$, i.e.,

$$\lim_{l \to \infty} P\left\{\sup_{i \geq l} |\xi_i - \xi_0| < \varepsilon\right\} = 1.$$

We shall denote convergence with probability 1 (almost surely) by $\xi \xrightarrow{\text{a.s.}} \xi_0$.

These two definitions reflect different requirements on the notion of convergence.

In the first case the event $\{|\xi_l - \xi_0| < \varepsilon\}$ selects a set of sequences for which the condition $|\xi_l - \xi_0| < \varepsilon$ is fulfilled for a given fixed l. Moreover, as l increases each particular sequence may or may not satisfy this condition. Convergence in probability is in a sense a "weak" convergence—it does not guarantee at all that each specific realization of $\xi_1, \ldots, \xi_l$ converges in the regular sense.

On the other hand, convergence with probability 1 is indeed a "strong" convergence. It implies that almost all realizations converge in the regular sense. Convergence almost surely may also be defined as follows:

Definition 2a. A sequence of random variables $\xi_1, \ldots, \xi_l, \ldots$ converges with probability 1 to ξ_0 if the measure of the set of realizations of the variables for which the limit

$$\lim_{l \to \infty} \xi_l = \xi_0$$

exists equals 1, i.e.,

$$P\left\{\lim_{l \to \infty} \xi_l = \xi_0\right\} = 1.$$

It is easy to verify that convergence with probability 1 implies convergence in probability. Indeed, since for any l the inequality

$$P\{|\xi_l - \xi_0| < \varepsilon\} \geq P\left\{\sup_{i \geq l} |\xi_i - \xi_0| < \varepsilon\right\}$$

is valid, the condition

$$\lim_{l \to \infty} P\left\{\sup_{i \geq l} |\xi_i - \xi_0| < \varepsilon\right\} = 1$$

implies

$$\lim_{l \to \infty} P\{|\xi_l - \xi_0| < \varepsilon\} = 1.$$

The converse is generally not true. Additional conditions under which convergence in probability implies convergence with probability 1 are given by the following lemma.

Lemma (Borel–Cantelli). *If for a random sequence $\xi_1, \ldots, \xi_l, \ldots$ there exists ξ_0 such that for any $\varepsilon > 0$ the inequality*

$$\sum_{i=1}^{\infty} P\{|\xi_i - \xi_0| \geq \varepsilon\} < \infty \tag{9.7}$$

is fulfilled, then the sequence $\xi_1, \ldots, \xi_l, \ldots$ converges to ξ_0 with probability 1.

PROOF. Denote by E_l^r the event that the inequality

$$|\xi_l - \xi_0| > \frac{1}{r} \qquad (r \text{ an integer})$$

is fulfilled. Consider the event S_l^r which consists in the occurrence of at least one of the events $E_l^r, E_{l+1}^r, \ldots, E_{l+i}^r, \ldots$:

$$S_l^r = \bigcup_{i=0}^{\infty} E_{l+i}^r.$$

We bound the probability of this event:

$$P\{S_l^r\} < \sum_{i=0}^{\infty} P\{E_{l+i}^r\} = \sum_{i=l+1}^{\infty} P\left\{|\xi_i - \xi_0| > \frac{1}{r}\right\}.$$

Since in view of the lemma's conditions the series (9.7) is convergent, we have

$$\lim_{l \to \infty} P\{S_l^r\} = 0. \tag{9.8}$$

Now consider the event S^r:

$$S^r = \bigcap_{l=1}^{\infty} S_l^r.$$

Since the event S^r implies any one of the events S_l^r, in view of (9.8) we have

$$P\{S^r\} = 0. \tag{9.9}$$

Finally set $S = \bigcup_{r=1}^{\infty} S^r$. It is easy to verify that the meaning of this event is as follows: there exists an r such that for each l ($l = 1, 2, \ldots$), for at least one i ($i = i(l)$) the inequality

$$|\xi_{l+i} - \xi_0| > \frac{1}{r}$$

is fulfilled. Since

$$P\{S\} \leq \sum_{r=1}^{\infty} P\{S^r\},$$

we have, in view of (9.9), that $P\{S\} = 0$, q.e.d. □

§3 Theorems on Interpreting Results of Indirect Experiments

Let A be a linear, completely continuous operator acting from the space L_2 into the space C, and let A^* be the conjugate for A. Then the operator A^*A is also a completely continuous operator. Let

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_m^2 \geq \cdots$$

be a complete system of its eigenvalues and

$$\varphi_1(t), \ldots, \varphi_m(t) \ldots \tag{9.10}$$

be a complete orthonormal system of its eigenfunctions.

Consider also the operator AA^*. It has the same set of eigenvalues, to which a complete orthonormal system of eigenfunctions

$$\psi_1(x), \ldots, \psi_m(x), \ldots \tag{9.11}$$

corresponds. Elements of (9.10) and (9.11) satisfy the relations

$$\left.\begin{aligned} A\varphi_p(t) &= \lambda_p \psi_p(x), \\ A^*\psi_p(x) &= \lambda_p \varphi_p(t), \end{aligned}\right\} \qquad p = 1, 2, \ldots.$$

A solution of the operator equation (9.1) can be expanded in a series in the system of functions (9.10):

$$f(t) = \sum_{p=1}^{\infty} \alpha_p^0 \varphi_p(t). \tag{9.12}$$

We shall consider the function

$$f_l(t, \alpha_{\text{emp}}) = \sum_{p=1}^{n(l)} \alpha_{\text{emp}}^p \varphi_p(t) \tag{9.13}$$

to be an approximation to the solution (9.12). Here $n(l)$ is an appropriate number of terms in the expansion (to be determined below) and $\alpha_{\text{emp}} = (\alpha_{\text{emp}}^1, \ldots, \alpha_{\text{emp}}^{n(l)})^{\mathrm{T}}$ is the vector of parameters which yields the minimum for the functional

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{j=1}^{l} \left(y_j - \sum_{p=1}^{n(l)} \lambda_p \alpha_p \psi_p(x_j) \right)^2. \tag{9.14}$$

It turns out that under certain assumptions concerning the solution (9.12) there exists a function $n(l)$ such that as the sample size increases the approximations obtained approach in probability the solution of the operator equation (9.1).

The following two theorems are valid.

Theorem 9.1. *Let a unique solution for the operator equation* (9.1) *exist. Then the sequence of approximations* $f_l(t, \alpha_{\text{emp}})$ *as* l *increases converges in probability to* $f(t)$ *in the metric* L_2, *provided only the function* $n(l)$ *satisfies*

$$n(l) \xrightarrow[l \to \infty]{} \infty, \tag{9.15}$$

$$\frac{1}{\lambda_{n(l)}^2} \sqrt{\frac{n(l) \ln l}{l}} \xrightarrow[l \to \infty]{} 0. \tag{9.16}$$

In addition we shall require that the action of the operator A^* be from L_2 into C.

Theorem 9.2. *Let a solution of the operator equation* (9.1) *be such that the conditions*

$$\sup_t \left| \sum_{p=m}^{\infty} \alpha_p^0 \varphi_p(t) \right| = T(m), \qquad T(m) \xrightarrow[m \to \infty]{} 0 \tag{9.17}$$

are fulfilled. Then the conditions (9.15) *and* (9.16) *are sufficient to assure convergence in probability of the functions* $f_l(t, \alpha_{\text{emp}})$ *to* $f(t)$ *in the* C *metric.*

Theorems 9.1 and 9.2 thus assert that if one approximates the solution of (9.1) by means of an expansion in a finite number of eigenfunctions of a self-adjoint operator A^*A, then under an appropriate choice of the number

of terms in the expansion (satisfying the conditions (9.15), (9.16)) the method of minimizing the empirical risk (9.14) assures the convergence in probability —as the sample size increases—of the solutions obtained to the desired one.

Below we shall show that under certain conditions the selection of $n(l)$ may be carried out using the minimization of the right-hand side of (8.69). It will thus be shown that the standard procedure of the method of structural risk minimization considered in Chapter 8 leads to the construction of a sequence of functions which converges to the solution of the operator equation (9.1).

We shall assume that the error ξ_i associated with the measurements of the function on the right-hand side of the operator equation (9.1) is defined by a probability density function $P(\xi)$ and satisfies the conditions $M\xi = 0$ and

$$\tau = \frac{\sqrt[p]{M\xi^{2p}}}{M\xi^2} < \infty, \qquad p > 2. \tag{9.18}$$

Furthermore, let the inequality

$$\sup_{\alpha} \frac{\sqrt[p]{M\left(y - \sum_{p=1}^{k} \lambda_p \alpha_p \psi_p(x)\right)^{2p}}}{M\left(y - \sum_{p=1}^{p} \lambda_p \alpha_p \psi_p(x)\right)^2} \leq \frac{\text{const}}{\lambda_k^2} = \tau_k \tag{9.19}$$

be fulfilled for some $p > 2$.

The inequality (9.19) follows from (9.18) provided

$$y_j = F(x_j, \alpha_0) + \xi_j = \sum_{p=1}^{k} \lambda_p \alpha_p^0 \psi_p(x_j) + \xi_j.$$

In this case

$$\left(y_j - \sum_{p=1}^{k} \lambda_p \alpha_p \psi_p(x_j)\right)^2 = \left(\xi_j - \sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x_j)\right)^2, \tag{9.20}$$

where

$$\beta_p = \alpha_p - \alpha_p^0.$$

It follows from (9.19) and (9.20) that

$$\tau_k = \sup_{\alpha} \frac{\sqrt[p]{M\left(\xi - \sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x)\right)^{2p}}}{M\left(\xi - \sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x)\right)^2}. \tag{9.21}$$

We shall bound separately the denominator and numerator on the right-hand side of (9.21):

$$M\left(\xi - \sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x)\right)^2 = \sigma^2 + \sum_{p=1}^{k} \lambda_p^2 \beta_p^2 = \sigma^2 + B, \tag{9.22}$$

where

$$B = \sum_{p=1}^{k} \lambda_p^2 \beta_p^2;$$

using the Minkowski inequality we obtain

$$\sqrt[p]{M\left(\xi - \sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x)\right)^{2p}} \leq 2\left(\sqrt[p]{M\xi^{2p}} + \sqrt[p]{M\left(\sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x)\right)^{2p}}\right) \tag{9.23}$$

The following bound is valid (as an operator acting from L_2 into C is bounded by L):

$$\sup_x \left(\sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x)\right)^2 \leq L \frac{B}{\lambda_k^2}, \tag{9.24}$$

Substituting (9.24) into (9.23) and taking into account

$$\frac{\sqrt[p]{M\xi^{2p}}}{M\xi^2} = \tau$$

we obtain

$$\sqrt[p]{M\left(\xi - \sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x)\right)^{2p}}$$

$$\leq 2\tau\sigma^2 + 2L\frac{B}{\lambda_k^2} \leq \frac{2R}{\hat{\lambda}_k^2}(\sigma^2 + B),$$

where $R = \max(\tau, L)$,

$$\hat{\lambda}_k \begin{cases} = 1, & \text{if } \lambda_k > 1, \\ = \lambda_k, & \text{if } \lambda_k \leq 1. \end{cases} \tag{9.25}$$

Substituting (9.22) and (9.25) into (9.21), we finally arrive at

$$\tau_k \leq \frac{2R}{\hat{\lambda}_k^2}.$$

In this case, according to Theorem 7.6, the inequality

$$I(\alpha) < \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau_n a(p)\sqrt{\dfrac{n\left(\ln\dfrac{2l}{n} + 1\right) - \ln\dfrac{\eta}{12}}{l}}}\right]_{\infty} \tag{9.26}$$

is fulfilled with probability $1 - \eta$ simultaneously for all functions expandable in terms of n ($n < l$) eigenvectors of the system (9.11). For each sample size l we shall use a number $n(l)$ of terms in the expansion such that firstly the restriction

$$n(l) < l^{1-\delta} \tag{9.27}$$

is fulfilled, where δ is an arbitrary small quantity, and furthermore the right-hand side (9.26) attains its minimum. (Here an additional condition appears

which requires that the number of terms in the expansion increase at a rate not exceeding $l^{1-\delta}$ as the sample size l increases.)

Using this modified version of defining the number of terms in the expansion, we satisfy the condition (9.16) stipulated in Theorems 9.1 and 9.2. In other words the following theorem is valid.

Theorem 9.3. *Let a solution of the operator equation* (9.1) *satisfy the condition*

$$\left\| \sum_{i=1}^{\infty} \alpha_p^0 \varphi_p(t) \right\|_{L_2} < \infty, \tag{9.28}$$

and let the conditions (9.19) *and* (9.27) *be fulfilled. Then using structural minimization of the bound* (9.26), *a number of terms in the expansion is determined such that with probability* 1 *the conditions*

(1) $$n(l) \xrightarrow[l\to\infty]{} \infty,$$

(2) $$\frac{1}{\lambda_{n(l)}^2} \sqrt{\frac{n(l) \ln l}{l}} \xrightarrow[l\to\infty]{} 0$$

are satisfied.

Thus Theorems 9.1 and 9.2 point to a class of algorithms which assure the convergence of the sequence of the functions obtained to the solution of the operator equation, while Theorem 9.3 asserts that the method of structural risk minimization determined by means of the bound (9.26) on a structure formed by a system of eigenfunctions belongs to this class.

In Section 6 examples are presented which show that the method of structural risk minimization for interpretation of results of indirect experiments is an efficient one. Here we would like to note that success in applying this method to ill-posed problems of interpreting measurements is probably due to the fact that for each finite l it determines a solution which possesses an extremal property (the image of a solution in E_2 yields a guaranteed minimum for the value of the expected risk), rather than to the fact that the sequence of solutions obtained converges to the desired solution of the operator equation (9.1).

§4 Proofs of the Theorems

We shall now prove the theorems stated in Section 3.

PROOF OF THEOREM 9.1. Let the conditions of the theorem be satisfied. Denote by

$$f_l(t, \alpha_{\text{emp}}) = \sum_{p=1}^{n(l)} \alpha_{\text{emp}}^p \varphi_p(t)$$

the preimage of the function

$$F_l(x, \alpha_{\text{emp}}) = \sum_{p=1}^{n(l)} \lambda_p \alpha_{\text{emp}}^p \psi_p(x)$$

which minimizes the value of the empirical risk

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{j=1}^{l} \left(y_j - \sum_{p=1}^{n(l)} \lambda_p \alpha_p \psi_p(x_j) \right)^2. \tag{9.29}$$

Our goal is to prove that $f_l(t, \alpha_{\text{emp}})$ converges in probability to $f(t)$ in the metric L_2, or equivalently that the sequence of random variables

$$v(l) = \int \left(\sum_{p=1}^{n(l)} \alpha_{\text{emp}}^p \varphi_p(t) - \sum_{p=1}^{\infty} \alpha_p^0 \varphi_p(t) \right)^2 dt \tag{9.30}$$

converges in probability to zero as l increases.

Note that

$$v(l) = \sum_{p=1}^{n(l)} \beta_p^2 + \sum_{p=n(l)+1}^{\infty} (\alpha_p^0)^2 = T_1(n(l)) + T_2(n(l)),$$

where $\beta_p = \alpha_{\text{emp}}^p - \alpha_p^0$.

Since the solution belongs to L_2, the sequence $T_2(n(l))$ tends to zero as $n(l)$ increases. Therefore, to prove the theorem it is sufficient to show that

$$T_1(n(l)) \xrightarrow[l \to \infty]{} 0.$$

We bound the quantity

$$T_1(n(l)) = \sum_{p=1}^{n(l)} \beta_p^2. \tag{9.31}$$

To do this we define a vector $\beta = (\beta_1, \ldots, \beta_{n(l)})^{\mathrm{T}}$ for which the minimum of the empirical risk is attained. We then rewrite (9.29) in the form

$$I_{\text{emp}}(\beta) = \frac{1}{l} \sum_{j=1}^{l} \hat{y}_j^2 - 2 \sum_{p=1}^{n(l)} \lambda_p \beta_p G_p + \sum_{p,q=1}^{n(l)} \lambda_p \beta_p \lambda_q \beta_q \sum_{j=1}^{l} \frac{\psi_p(x_j) \psi_q(x_j)}{l}, \tag{9.32}$$

where

$$G_p = \frac{1}{l} \sum_{j=1}^{l} \hat{y}_j \psi_p(x_j), \qquad \hat{y}_j = \xi_j + \sum_{p=n+1}^{\infty} \lambda_p \alpha_p^0 \psi_p(x_j).$$

Denote by $\|K\|$ the covariance matrix with elements K_{pq} given by

$$K_{pq} = \frac{1}{l} \sum_{i=1}^{l} \psi_p(x_i) \psi_q(x_i),$$

and by G the n-dimensional vector with coordinates $G_1, \ldots, G_n$. Then the vector $\gamma = (\beta_1\lambda_1, \ldots, \beta_n\lambda_n)^T$ which yields the minimum for (9.32) is given by

$$\gamma = \|K\|^{-1}G.$$

Therefore the bound

$$|\gamma|^2 = |\|K\|^{-1}G|^2 \leq |\|K\|^{-1}|^2|G|^2 \tag{9.33}$$

is valid. On the other hand the inequality

$$|\gamma|^2 = \sum_{p=1}^{n(l)} (\beta_p\lambda_p)^2 > \lambda_{n(l)}^2 \sum_{p=1}^{n(l)} \beta_p^2 = \lambda_{n(l)}^2 T_1(n(l)) \tag{9.34}$$

holds. From the inequalities (9.33) and (9.34) we obtain

$$T_1(n(l)) < \frac{1}{\lambda_{n(l)}^2} |\|K\|^{-1}|^2|G|^2. \tag{9.35}$$

Thus to prove the theorem it is sufficient to bound from above the norm of the matrix $\|K\|^{-1}$ and the norm of the vector G.

We now bound $\|K\|^{-1}$. We note that the norm of the matrix $\|K\|$ does not exceed $\mu_{\max}$, the largest eigenvalue of the matrix, and the norm of the matrix $\|K\|^{-1}$ does not exceed $1/\mu_{\min}$, where $\mu_{\min}$ is the smallest eigenvalue of the matrix $\|K\|$.

We now bound $\mu_{\min}$ from below. For this purpose consider a positive definite quadratic form

$$F_n(x, \gamma) = \left(\sum_{p=1}^{n} \gamma_p \psi_p(x) \right)^2,$$

which we shall examine in the domain $\sum_{p=1}^{n} \gamma_p^2 \leq 1$. Since a completely continuous operator A acting from L_2 into C is bounded, $\|A\| < L$, the inequality

$$\sup_x \left| \sum_{p=1}^{n} \lambda_p \gamma_p \psi_p(x) \right| \leq \|A\| \left\| \sum_{p=1}^{n} \gamma_p \varphi_p(t) \right\| < L \sqrt{\sum_{p=1}^{n} \gamma_p^2}$$

holds, which implies that in the domain $\sum_{p=1}^{n} \gamma_p^2 \leq 1$ the inequality

$$\sup_{x,\gamma} \left| \sum_{p=1}^{n} \gamma_p \psi_p(x) \right| < L \sqrt{\sum_{p=1}^{n} \frac{\gamma_p^2}{\lambda_p^2}} \leq \frac{L}{\lambda_n}$$

is satisfied and hence

$$\sup_x F_n(x, \gamma) < \frac{L^2}{\lambda_n^2}. \tag{9.36}$$

Now consider the expression

$$\frac{1}{l} \sum_{i=1}^{l} F_n(x_i, \gamma) = \frac{1}{l} \sum_{i=1}^{l} \left(\sum_{p=1}^{n} \gamma_p \psi_p(x_i) \right)^2.$$

Observe that

$$MF_n(x, \gamma) = \sum_{p=1}^{n} \gamma_p^2,$$
$$\frac{1}{l} \sum_{i=1}^{l} F_n(x_i, \gamma) = \sum_{p,q=1}^{n} \gamma_p \gamma_q K_{pq}. \tag{9.37}$$

Using a rotation transformation, we arrive at a new, twice orthogonal system of functions $\psi_1'(x), \ldots, \psi_n'(x)$ such that

$$MF_n(x, \gamma') = \sum_{p=1}^{n} (\gamma_p')^2,$$
$$\frac{1}{l} \sum_{i=1}^{l} F_n(x_i, \gamma') = \sum_{p=1}^{n} \mu_p (\gamma_p')^2, \tag{9.38}$$

where $\mu_1, \ldots, \mu_n$ are the eigenvalues of the matrix $\|K\|$.

To bound the eigenvalues we utilize the theorem on the uniform convergence of the means to their mathematical expectations for a class of bounded functions (Theorem 7.3). Since the functions $F(x, \gamma')$ for $\|\gamma'\| \leq 1$ are bounded by the quantity L^2/λ_n^2, the inequality

$$P\left\{\sup_{\gamma'} \left| MF_n(x, \gamma') - \frac{1}{l} \sum_{i=1}^{l} F_n(x_j, \gamma') \right| > \varkappa \frac{L^2}{\lambda_n^2}\right\} < 9 \frac{(2l)^n}{n!} e^{-\varkappa^2 l/4}$$

is valid (cf. Section 3 of Chapter 7); taking (9.38) into account, we obtain

$$P\left\{\sup_{\gamma'} \left| \sum_{p=1}^{n} (\gamma_p')^2 (1 - \mu_p) \right| > \varkappa \frac{L^2}{\lambda_n^2}\right\} < 9 \frac{(2l)^n}{n!} e^{-\varkappa^2 l/4}. \tag{9.39}$$

We shall require that the probability not exceed $9/\ln l$. For this purpose it is sufficient for $\varkappa$ to be at least $\varkappa^*$, where

$$\varkappa^* = \frac{2L^2}{\lambda_n^2} \sqrt{\frac{n\left(\ln \frac{2l}{n} + 1\right) + \ln \ln l}{l}}. \tag{9.40}$$

It follows from (9.39) and (9.40) that with probability $1 - (9/\ln l)$ all the eigenvalues $\mu_1, \ldots, \mu_n$ are located in the interval

$$1 - \varkappa^* \leq \mu_i \leq 1 + \varkappa^*; \tag{9.41}$$

this implies that with probability $1 - (9/\ln l)$ the inequality

$$\mu > 1 - \varkappa^* \tag{9.42}$$

is fulfilled.

Substituting (9.42) into (9.35), we obtain that the inequality

$$T_1(n) < \frac{|G|^2}{\lambda_n^2\left(1 - \frac{2L^2}{\lambda_n^2}\sqrt{\frac{n\left(\ln\frac{2l}{n} + 1\right) + \ln\ln l}{l}}\right)^2} \tag{9.43}$$

is valid with probability $1 - (9/\ln l)$. It remains to bound the quantity $|G|^2$:

$$|G|^2 = \sum_{p=1}^{n} G_p^2 = \sum_{p=1}^{n} \frac{1}{l^2}\left(\sum_{i=1}^{l} \hat{y}_i \psi_p(x_i)\right)^2.$$

For this purpose we compute the mathematical expectation

$$M|G|^2 = M\sum_{p=1}^{n} G_p^2 \leq \frac{\sigma^2 + \|A\|^2 T_2(0)}{l}\, n = R \cdot \frac{n}{l},$$

where T and R are constants which do not depend on l and n. To bound $|G|$ we utilize Chebyshev's inequality for the first moment of a positive random variable ξ:

$$P\{\xi > \varepsilon\} < \frac{M\xi}{\varepsilon},$$

where $\varepsilon = (Rn \ln l)/l$. Since $M|G|^2 < Rn/l$, we obtain

$$P\left\{|G|^2 > \frac{Rn \ln l}{l}\right\} < \frac{1}{\ln l}.$$

Thus with probability $1 - 1/\ln l$,

$$|G|^2 \leq \frac{Rn \ln l}{l}. \tag{9.44}$$

Substituting (9.44) into (9.43), we obtain that for l sufficiently large the inequality

$$T_1(n) < c\, \frac{n \ln l}{l\lambda_n^2\left(1 - \frac{c}{\lambda_n^2}\sqrt{\frac{n \ln l}{l}}\right)^2} \tag{9.45}$$

is fulfilled with probability $1 - (10/\ln l)$, where c is a constant.

The inequality (9.45) implies that $T_1(n(l))$ tends to zero in probability as

$$\frac{1}{\lambda_{n(l)}^2}\sqrt{\frac{n(l) \ln l}{l}} \xrightarrow[l\to\infty]{} 0.$$

The theorem is proved. □

PROOF OF THEOREM 9.2. Now let the solution of the operator equation (9.1) obey the additional restriction

$$\sup_t \left| \sum_{p=n}^{\infty} \alpha_p \varphi_p(t) \right| \xrightarrow[n\to\infty]{} 0. \tag{9.46}$$

We shall show that in this case the conditions

$$\begin{gathered} n(l) \xrightarrow[l\to\infty]{} \infty, \\ \frac{1}{\lambda^2_{n(l)}} \sqrt{\frac{n(l) \ln l}{l}} \xrightarrow[l\to\infty]{} 0 \end{gathered} \tag{9.47}$$

are sufficient in order that the sequence of solutions $f_l(t, \alpha_{\text{emp}})$ converge in probability to the solution of the operator equation (9.1) in the metric C. We use the notation

$$v(l) = \sup_t \left| \sum_{p=1}^{\infty} \alpha_p^0 \varphi_p(t) - \sum_{p=1}^{n(l)} \alpha_{\text{emp}}^p \varphi_p(t) \right|,$$

where $\alpha_{\text{emp}} = (\alpha_{\text{emp}}^1, \ldots, \alpha_{\text{emp}}^{n(l)})^{\text{T}}$ is the vector which yields the minimal value for (9.29). Our purpose is to prove that

$$v(l) \xrightarrow[l\to\infty]{p} 0.$$

Observe that

$$v(l) \le \sup_t \left| \sum_{p=1}^{n(l)} \beta_p \varphi_p(t) \right| + \sup_t \left| \sum_{p=n(l)+1}^{\infty} \alpha_p^0 \varphi_p(t) \right|, \tag{9.48}$$

where $\beta_p = \alpha_p^0 - \alpha_{\text{emp}}^p$.

In view of the condition (9.46) of the theorem, the second summand in (9.48) tends to zero with increase in l. It is therefore sufficient to verify that

$$T_3(n(l)) = \sup_t \left| \sum_{p=1}^{n(l)} \beta_p \varphi_p(t) \right| \xrightarrow[l\to\infty]{} 0. \tag{9.49}$$

To prove this we shall use the bound

$$T_3^2(n(l)) < \frac{\text{const}}{\lambda_n^2} \sum_{p=1}^{n} \beta_p^2, \tag{9.50}$$

which is valid because the operator A^* is bounded.

In the course of the proof of Theorem 9.1 it was shown that the bound

$$T_1(n) = \sum_{p=1}^{n} \beta_p^2 < \frac{\text{const } n \ln l}{l\lambda_n^2 \left(1 - \dfrac{\text{const}}{\lambda_n^2} \sqrt{\dfrac{n \ln l}{l}}\right)^2}$$

holds with probability $1 - (10/\ln l)$. Substituting this bound into (9.50), we obtain that with probability $1 - (10/\ln l)$ the inequality

$$T_3^2(n(l)) < \frac{\dfrac{\text{const}}{\lambda_n^4}\dfrac{n \ln l}{l}}{\left(1 - \dfrac{\text{const}}{\lambda_n^2}\sqrt{\dfrac{n \ln l}{l}}\right)^2} \tag{9.51}$$

is satisfied. The bound (9.51) implies that $T_3^2(n)$ approaches zero in probability provided that

$$\frac{1}{\lambda_{n(l)}^2}\sqrt{\frac{n(l) \ln l}{l}} \xrightarrow[l \to \infty]{} 0.$$

Theorem 9.2 is thus proved. □

PROOF OF THEOREM 9.3. Let the number $n(l)$ of terms in the expansion of the solution of the operator equation be determined by the minimal value of the bound (9.26). We shall show that if a solution of the operator equation $f(t)$ satisfies

$$\left\| \sum_{p=1}^{\infty} \alpha_p^0 \varphi_0(t) \right\|_{L_2} < \infty, \tag{9.52}$$

then the algorithm under consideration for determining the number of terms in the expansion satisfies

$$n(l) \xrightarrow[l \to \infty]{} \infty, \tag{9.53}$$

$$\frac{1}{\lambda_n^2}\sqrt{\frac{n(l) \ln l}{l}} \xrightarrow[l \to \infty]{} 0. \tag{9.54}$$

First we verify that (9.53) is valid. Assume the contrary. Let $\alpha_n^0 \neq 0$, $r < n$, and also let the inequality

$$\frac{\dfrac{1}{l}\sum_{i=1}^{l}\left(y_i - \sum_{p=1}^{r}\lambda_p \alpha_{\text{emp}}^p \psi_p(x_i)\right)^2}{1 - \dfrac{c}{\lambda_r^2}\sqrt{\dfrac{r\left(\ln \dfrac{2l}{r} + 1\right) - \ln \dfrac{\eta}{12}}{l}}} < \frac{\dfrac{1}{l}\sum_{i=1}^{l}\left(y_i - \sum_{p=1}^{n}\lambda_p \alpha_{\text{emp}}^p \psi_p(x_i)\right)^2}{1 - \dfrac{c}{\lambda_n^2}\sqrt{\dfrac{n\left(\ln \dfrac{2l}{n} + 1\right) - \ln \dfrac{\eta}{12}}{l}}} \tag{9.55}$$

be fulfilled for any $l > l_0$. In view of Theorem 7.6 the inequality

$$I(\alpha_{\text{emp}}, r) < \frac{\dfrac{1}{l}\sum_{i=1}^{l}\left(y_i - \sum_{p=1}^{n}\lambda_p \alpha_{\text{emp}}^p \psi_p(x_i)\right)^2}{1 - \dfrac{c}{\lambda_n^2}\sqrt{\dfrac{n\left(\ln \dfrac{2l}{n} + 1\right) - \ln \dfrac{\eta}{12}}{l}}}$$

is valid with probability $1 - \eta$ for l sufficiently large. Represent the quantity $I(\alpha_{\text{emp}}, r)$ in the form

$$I(\alpha_{\text{emp}}, r) = M\left(y - \sum_{p=1}^{r} \lambda_p \alpha_{\text{emp}}^p \psi_p(x)\right)^2$$

$$= M\left(\xi + \Delta(x, r) - \sum_{p=1}^{r} \beta_p \lambda_p \psi_p(x)\right)^2,$$

where

$$\Delta(x, r) = \sum_{p=r+1}^{\infty} \lambda_p \alpha_p^0 \psi_p(x), \qquad \beta_p = \alpha_{\text{emp}}^p - \alpha_p^0,$$

and bound this quantity from below:

$$I(\alpha_{\text{emp}}, r) > I(\alpha_0, r) = \sigma^2 + M\Delta^2(x, r) \geq \sigma^2 + \sum_{p=r+1}^{\infty} (\alpha_p^0 \lambda_p)^2.$$

Thus the bound

$$\sigma^2 + \sum_{p=r+1}^{\infty} (\alpha_p^0 \lambda_p)^2 < \frac{\frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{p=1}^{n} \alpha_{\text{emp}}^p \lambda_p \psi_p(x_i)\right)^2}{1 - \frac{c}{\lambda_n^2} \sqrt{\frac{n\left(\ln \frac{2l}{n} + 1\right) - \ln \frac{\eta}{12}}{l}}} \tag{9.56}$$

is valid with probability $1 - \eta$. We now transform and bound the expression appearing in the numerator on the right-hand side of (9.56):

$$I_{\text{emp}}(\alpha_{\text{emp}}, n) = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{p=1}^{n} \lambda_p \alpha_{\text{emp}}^p \psi_p(x_i)\right)^2$$

$$= \frac{1}{l} \sum_{i=1}^{l} \left(\xi_i + \Delta(x_i, n) - \sum_{p=1}^{n} \lambda_p \beta_p \psi_p(x_i)\right)^2$$

$$\leq \frac{1}{l} \sum_{i=1}^{l} (\xi_i + \Delta(x_i, n))^2$$

$$= \frac{1}{l} \sum_{i=1}^{l} \xi_i^2 + \frac{1}{l} \sum_{i=1}^{l} \Delta^2(x_i, n) + \frac{2}{l} \sum_{i=1}^{l} \xi_i \Delta(x_i, n).$$

Note that in view of the law of large numbers

$$\frac{1}{l} \sum_{i=1}^{l} \xi_i^2 \xrightarrow[l \to \infty]{} \sigma^2,$$

$$\frac{1}{l} \sum_{i=1}^{l} \xi_i \Delta(x_i, n) \xrightarrow[l \to \infty]{} 0.$$

$$\frac{1}{l} \sum_{i=1}^{l} \Delta^2(x_i, n) \xrightarrow[l \to \infty]{} \sum_{p=n+1}^{\infty} (\lambda_p \alpha_p^0)^2,$$

Therefore the inequality

$$\sigma^2 + \sum_{p=r+1}^{\infty} (\lambda_p \alpha_p^0)^2 < \sigma^2 + \sum_{p=n+1}^{\infty} (\lambda_p \alpha_p^0)^2 \tag{9.57}$$

is satisfied with probability $1 - \eta$ for l sufficiently large. However, for $r < n$ the inequality (9.57) is obviously invalid with probability 1. The contradiction obtained proves the validity of (9.53).

We now show that (9.54) is also valid. For this purpose note that the inequalities

$$\frac{1}{l} \sum_{i=1}^{l} (\xi_i + \Delta(x_i, n))^2 > \min_{\alpha} I_{\text{emp}}(\alpha, n)$$

$$> \min_{\alpha, \gamma} \frac{1}{l} \sum_{i=1}^{l} \left(\xi_i - \sum_{p=1}^{n} \lambda_p \alpha_p \psi_p(x_i) - \gamma \Delta(x_i, n) \right)^2 \tag{9.58}$$

always hold. Compute the mathematical expectation of the left-hand side of the inequality (9.58):

$$M\left\{ \frac{1}{l} \sum_{i=1}^{l} (\xi_i + \Delta(x_i, n))^2 \right\} = \sigma^2 + \sum_{p=r+1}^{\infty} (\alpha_p \lambda_p)^2 = \sigma^2 + T_1(n).$$

Observe that for a fixed n the relation

$$\frac{1}{\lambda_n^2} \sqrt{\frac{n \ln l}{l}} \xrightarrow[l \to \infty]{} 0$$

is valid. Therefore the inequality

$$\lim_{l \to \infty} \frac{I_{\text{emp}}(\alpha_{\text{emp}}, r)}{1 - \dfrac{\text{const}}{\lambda_r^2} \sqrt{\dfrac{r\left(\ln \dfrac{2l}{r}\right) - \ln \dfrac{\eta}{12}}{l}}} < \sigma^2 + T_1(r) \tag{9.59}$$

is fulfilled. Since the inequality (9.59) is valid for any r and the conditions $T_1(r) \to_{r \to \infty} 0$ and $n(l) \to_{l \to \infty} \infty$ are fulfilled, it follows that the inequality

$$\lim_{l \to \infty} \min_{r < l^{1-\delta}} \frac{I_{\text{emp}}(\alpha_{\text{emp}}, r)}{1 - \dfrac{\text{const}}{\lambda_r^2} \sqrt{\dfrac{r \ln l}{l}}} \leq \sigma^2$$

holds. On the other hand we utilize the following bounds on the mean and the variance:

$$MI_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, r) = M\frac{1}{l}\sum_{i=1}^{l}\left(\xi_i - \sum_{p=1}^{l}\lambda_0\alpha_{\text{emp}}^p\psi_p(x_i) - \gamma_{\text{emp}}\Delta(x_i, r)\right)^2$$

$$= \sigma^2\left(1 - \frac{r+1}{l}\right), \tag{9.60}$$

$$D[I_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, r)] = M(I_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, r) - MI_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, r))^2$$

$$< \frac{M\xi^4 + \sigma^4}{l^2}(r+1) = R\frac{r+1}{l^2}. \tag{9.61}$$

Here α_{emp} and γ_{emp} are the values of the parameters which minimize $I_{\text{emp}}(\alpha, \gamma, r)$. (We shall verify these bounds below.)

Now use Chebyshev's inequality and obtain

$$P\left\{\left|I_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}; n(l)) - \sigma^2\left(1 - \frac{n(l)+1}{l^2}\right)\right| > \varepsilon\right\} < \frac{R}{l^2\varepsilon^2}(n(l)+1).$$

According to the condition of the theorem, $n(l) < l^{1-\delta}$. Therefore

$$\sum_{l=1}^{\infty} P\left\{\left|\sigma^2\left(1 - \frac{n(l)+1}{l}\right) - I_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, n(l))\right| > \varepsilon\right\}$$

$$< R\sum_{l}^{\infty}\frac{l^{1-\delta}+1}{l^2\varepsilon^2} < \infty,$$

and consequently the convergence

$$\lim_{l\to\infty} I_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, n(l)) = \sigma^2$$

is valid with probability 1 according to the Borel–Cantelli lemma (see Section 2). Therefore with probability 1 the inequality

$$\lim_{l\to\infty}\min_{n(l)}\frac{I_{\text{emp}}(\alpha_{\text{emp}}, n(l))}{1 - \dfrac{\text{const}}{\lambda_{n(l)}^2}\sqrt{\dfrac{n(l)\ln l}{l}}} \le \sigma^2$$

as well as the equality

$$\lim_{l\to\infty} I_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, n(l)) = \sigma^2$$

are fulfilled. This implies that with probability 1

$$\lim_{l\to\infty}\frac{\text{const}}{\lambda_{n(l)}^2}\sqrt{\frac{n(l)\ln l}{l}} = 0. \tag{9.62}$$

This expression constitutes the statement of Theorem 9.3.

In the course of the proof we have used the equality (9.60) and the inequality (9.61). We shall now derive them:

$$MI_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, r) = M\frac{1}{l}\sum_{i=1}^{l}\left(\xi_i - \sum_{p=1}^{r}\lambda_p\alpha_{\text{emp}}^p\psi_p(x_i) - \gamma_{\text{emp}}\Delta(x, r)\right)^2.$$

Using a rotation transformation we arrive at a coordinate system $\psi'_1(x), \ldots, \psi'_r(x), \psi'_{r+1}(x)$ such that

$$\frac{1}{l}\sum_{i=1}^{l}\psi'_p(x_i)\psi'_q(x_i) = \begin{cases}\mu_p & \text{for } p = q,\\ 0 & \text{for } p \neq q.\end{cases}$$

In this coordinate system

$$I_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, r) = \frac{1}{l}\sum_{i=1}^{l}\xi_i^2 - \sum_{p=1}^{r+1}\frac{G_p^2}{\mu_p},$$

where

$$G_p = \frac{1}{l}\sum_{i=1}^{l}\xi_i\psi'_p(x_i).$$

We have thus obtained

$$MI_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, r) = \sigma^2 - \sum_{p=1}^{r+1}M\sum_{i,j=1}^{l}\frac{\xi_i\xi_j\psi'_p(x_i)\psi'_p(x_j)}{l^2\mu_p} = \sigma^2\left(1 - \frac{r+1}{l}\right),$$

$$D[I_{\text{emp}}(\alpha_{\text{emp}}, \gamma_{\text{emp}}, r)] = \sum_{p=1}^{r+1}\left[M\left(\frac{G_p^2}{\mu_p}\right)^2 - \left(M\frac{G_p^2}{\mu_p}\right)^2\right] \leq \frac{r+1}{l^2}R.$$

The theorem is proved. □

§5 Methods of Polynomial and Piecewise Polynomial Approximations

We have seen that using the method of structural risk minimization one can obtain a sequence of approximations which converges as the number of observations increases to the desired solution of the operator equation. However, the convergence is assured only under the condition that approximations are chosen to be in the form of expansions in terms of eigenfunctions of the operator A^*A. It is not always simple to obtain eigenfunctions of the operator A^*A. It would therefore be desirable to replace the expansion of the solution in terms of eigenfunctions of an operator by an expansion in terms of some other system of functions.

In this section we shall consider two types of approximations—polynomial and piecewise polynomial.

The basic property of polynomial approximations, stated in Weierstrass's theorem, asserts that any continuous function on the interval $[a, b]$ may be approximated in a uniform metric with arbitrary precision by a polynomial. In this book we choose, as an approximation to the function $y(x)$, a function which minimizes in $F(x, \alpha)$ the functional

$$I(\alpha) = \int (y(x) - F(x, \alpha))^2 \, dx. \tag{9.63}$$

The question arises: given an arbitrary continuous function $y(x)$, does the sequence

$$F(x, \alpha_1^0), \ldots, F(x, \alpha_r^0), \ldots \tag{9.64}$$

of polynomials of degrees $r = 0, 1, 2, \ldots$, each one of which yields the minimum for (9.63) in the class of polynomials of the corresponding degree, converge to $y(x)$ in a uniform metric?

The answer is no. The Lozinskiĭ - Kharshiladze theorem [30] asserts that there exists a continuous function $y(x)$ to which the sequence (9.64) does not converge uniformly.

Thus the idea of minimizing the mean squared deviation in order to obtain a uniform polynomial approximation for a continuous function is unacceptable. This immediately implies that one cannot hope to obtain a uniform approximation to a regression by minimizing the expected risk if the approximation is carried out in the class of polynomials.

The possibility of constructing uniform approximations to regressions using the method of expected-risk minimization is connected with a *piecewise polynomial approximation*, i.e., the so-called *spline approximations*.

Consider piecewise polynomial approximations of a function on the interval $[a, b]$. We subdivide the interval $[a, b]$ into N parts using the points $a = a_0, a_1, \ldots, a_{N+1} = b$. On each interval $[a_i, a_{i+1}]$ we shall approximate the function $y(x)$ by means of a polynomial of fixed degree m. Thus the function is approximated by means of $N + 1$ pieces of polynomials (each one on its own interval). Polynomials are chosen in such a manner that at points $a_1, \ldots, a_N$ the approximation obtained will be continuous together with its first $m - 1$ derivatives. Such a piecewise polynomial approximation is called a spline of degree m conjugated on the grid $(a_1, \ldots, a_N)$. We shall assume that the points of conjugation are fixed and are defined uniformly on $[a, b]$ $(a_i = a_0 + \Delta i, \Delta = (b - a)/N)$.

Denote by $V_N^m(x, \alpha)$ the class of splines of degree m with N conjugations defined on a uniform grid, and by $V_N^m(x, \alpha_{\text{emp}})$ the spline which yields the minimal value for the emprical functional

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - V_N^m(x_i \alpha))^2. \tag{9.65}$$

Now let a relation be defined which associates the number of conjugations N with the sample size l, i.e., $N = N(l)$.

Consider a sequence of splines

$$V^m_{N(1)}(x, \alpha_{\text{emp}}), \ldots, V^m_{N(l)}(x, \alpha_{\text{emp}}) \ldots \tag{9.66}$$

of degree m, possessing $N(1), \ldots, N(l), \ldots$ conjugations and minimizing the empirical risk on a sample $i = 1, 2, \ldots, l, \ldots$ (the samples are formed randomly and independently according to the density $P(x, y) = P(y|x)P(x)$).

The following theorem is valid.

Theorem 9.4 (Mihal'skiĭ). *Let the regression be determined by a continuous function $y(x)$. Then the sequence* (9.66) *converges with probability* 1 *in a uniform metric to the regression $y(x)$ provided only that the density $P(x)$ is absolutely continuous with respect to the uniform density and the conditions*

$$N(l) \xrightarrow[l \to \infty]{} \infty,$$

$$\frac{N^4(l) \ln l}{l} \xrightarrow[l \to \infty]{} 0$$

are fulfilled.

If, moreover, the stronger conditions

$$\frac{N^{2(2+p)}(l) \ln l}{l} \xrightarrow[l \to \infty]{} 0 \tag{9.67}$$

are fulfilled and the regression $y(x)$ is continuous together with its p derivatives, then the sequence

$$[V^m_{N(1)}(x, \alpha_{\text{emp}})]^{(p)}, \ldots, [V^m_{N(l)}(x, \alpha_{\text{emp}})]^{(p)}, \ldots,$$

constructed from the pth *derivatives of the splines* (9.66), *converges with probability* 1 *in a uniform metric to the function $y^{(p)}(x)$ which is the* pth *derivative of the regression.*

Remark. The theorem implies that the condition (9.67) guarantees estimation in the class of splines of the pth continuous derivative of a function $F(x)$ based on the values of this function measured at l randomly chosen points (l a sufficiently large number), i.e., it guarantees an approximate solution of the integral equation

$$\int_a^b \frac{(x-t)^{p-1}}{(p-1)!} \theta(x-t) f(t)\, dt = F(x) - \sum_{k=0}^{p-1} \frac{F^{(k)}(a)}{k!}$$

based on observations $y_i = F(x_i) + \xi_i$ $(i = 1, 2, \ldots, l)$.

Below, when interpreting results of experiments, we shall seek solutions in the form of expansions in terms of splines.

§6 Methods for Solving Ill-posed Measurement Problems

In this section we shall present examples of the use of the method of ordered risk minimization to estimate solutions of the linear operator equation

$$Af(t) = F(x) \tag{9.68}$$

from empirical data $x_1, y_1; \ldots; x_l, y_l$ ($y_i = F(x_i) + \xi_i$, x a random variable distributed according to the uniform distribution on $[a, b]$). The estimation is carried out in the class of splines.

It will be shown in Addendum 2 that any spline $V_N^m(t, \alpha)$ of order m with N conjugations can be represented as a linear combination of a system of $N + m + 1$ canonical splines of degree m with N conjugations,

$$\pi_1(t), \ldots, \pi_{N+m+1}(t). \tag{9.69}$$

In other words, the equality

$$V_N^m(t, \alpha) = \sum_{i=1}^{N+m+1} \alpha_i \pi_i(t)$$

is valid, where $\alpha = (\alpha_1, \ldots, \alpha_{N+m+1})$ are coefficients which define specific piecewise polynomial approximations in the class of splines of degree m with N conjugations.

When constructing a spline approximation to a solution of Equation (9.68), the problem is firstly to determine an appropriate number N of points of conjugation of the spline, and secondly to identify the coefficients $\alpha_1, \ldots, \alpha_{N+m+1}$ in the expansion.

Consider the images of the canonical system (9.69) in E_2,

$$\mu_1(x) = A\pi_1(t), \ldots, \qquad \mu_{N+m+1}(x) = A\pi_{N+m+1}(t),$$

and choose for a solution of the operator equation (9.68) a spline $V_N^m(t, \alpha^*)$ such that its image $F(x, \alpha^*)$ guarantees a small value of the risk:

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(y|x)\, dy\, dx.$$

According to Theorem 7.6 the inequality

$$I(\alpha) < \left[\frac{\frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{j=1}^{N+m+1} \alpha_j \mu_j(x_i) \right)^2}{1 - 2\tau_N a(p) \sqrt{\frac{(N+m+1)\left(\ln \frac{2l}{N+m+1} + 1 \right) - \ln \frac{\eta}{12}}{l}}} \right]_+$$

is fulfilled with probability $1 - \eta$ simultaneously for all splines with N conjugates. As a solution of the operator equation we choose a spline function (i.e., the number of conjugates N and a value of α) that yields the minimum for the right-hand side of this inequality. Although the convergence of approximations obtained using structural minimization of the risk to the

solution of an operator equation was proved only for expansions in terms of eigenfunctions, examples of successful solution of practical problems of interpreting results of indirect experiments in the class of splines permit us to recommend this expansion for solving Fredholm's integral equations of type I as well.

EXAMPLE 1 (A Problem in Nuclear Spectroscopy). At the input of a measuring device energy enters whose frequency is distributed according to $f(t)$ (t is the frequency). At the output of the device an experimental spectrum $F(x)$ is observed. The relation between the input and the output is given by equation

$$\int_a^b \left[1 - \frac{t}{x}\right]_+ f(t)\,dt = F(x),$$

where a and b are the endpoints of the emitted spectrum, and

$$[z]_+ = \begin{cases} z & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

Based on the observations, it is required to estimate $f(t)$.

In Figure 8 measurements of function $F(x)$ are shown (only each second observation is indicated). A total of 40 measurements were carried out. The measurements were subject to a uniformly distributed error concentrated on the interval $[-c, c]$. The value of c was chosen to be 2% of the maximum of $F(x)$.

The true spectrum (bold line) and the spline approximation obtained using the method of structural risk minimization are presented in Figure 9.

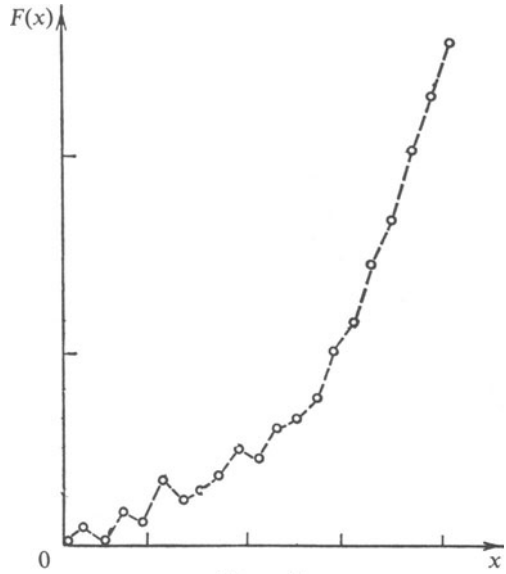

Figure 8

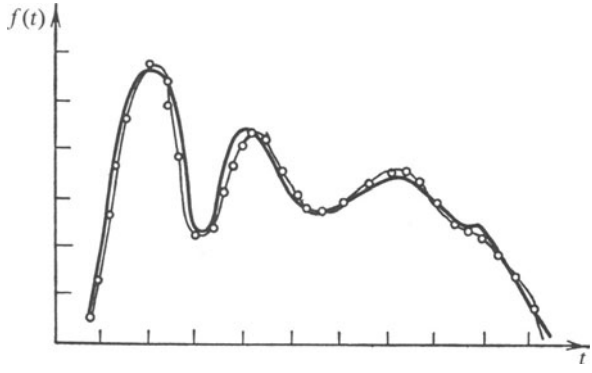

Figure 9

EXAMPLE 2 (The Inverse Problem of Gravimetry). The integral equation

$$\frac{2}{(\rho_1 - \rho_2)\pi} \int_a^b \frac{Hf(t)}{H^2 + (x - t)^2}\, dt = F(x)$$

describes the anomaly of the force of gravity on the surface of the earth created by a mass of density ρ_1 separated from the surrounding medium with density ρ_2 by the boundary $f(t)$; here H is the depth of the bedding of the mass which causes the anomaly. Based on the measurements of the anomaly $F(t)$, it is required to estimate the boundary $f(t)$.

The actual function (bold line) and the spline-approximated solution obtained by the method of structural risk minimization (thin line) are presented in Figure 10. The solution was obtained from 40 measurements conducted with uniform error whose amplitude was 12% of the maximum of $F(x)$.

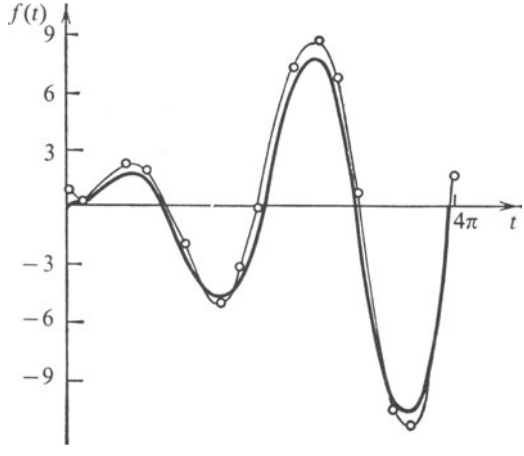

Figure 10

EXAMPLE 3 (A Problem in Estimating Derivatives). The problem of estimating the nth derivative in the class of continuous functions can be reduced to a solution of the following integral equation:

$$\int_a^b \frac{[x-t]_+^{n-1}}{(n-1)!} f(t)\,dt = F(x) - \sum_{j=0}^{n-1} \frac{F^{(j)}(a)}{j!}.$$

Solutions of this problem are presented in Figures 11–14 for $n = 1, 2, 3$ in the case when the function $F(x)$ was measured at 40 points. The function $F(t)$ (bold line) and its measurements are presented in Figure 11 (only every second observation of the function is indicated). The measurements were carried out subject to an error distributed uniformly with amplitude equal to 5% of the maximum of $F(x)$.

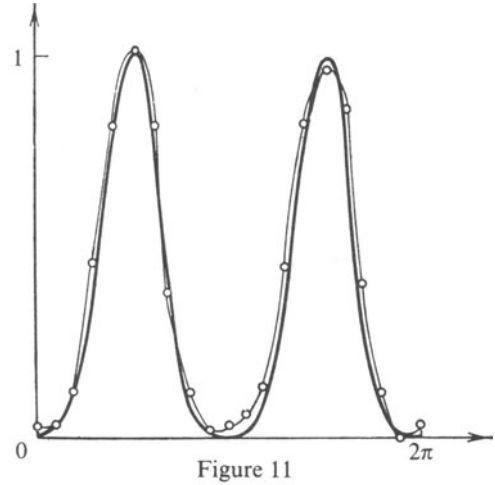

Figure 11

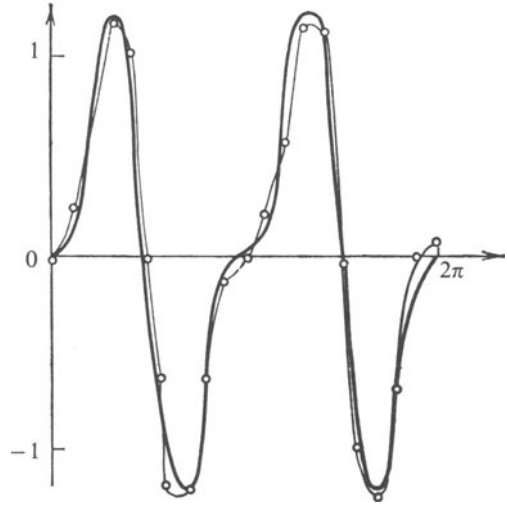

Figure 12

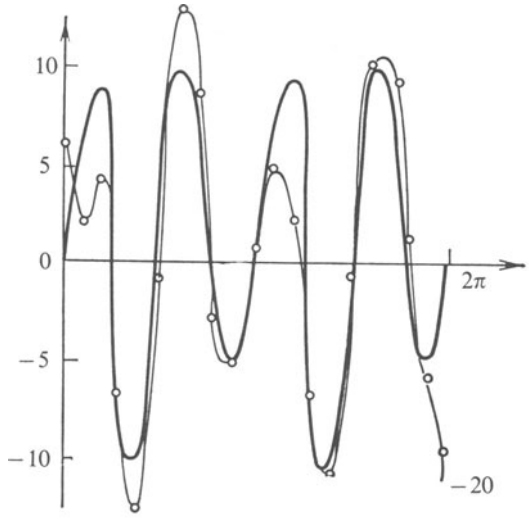

Figure 13

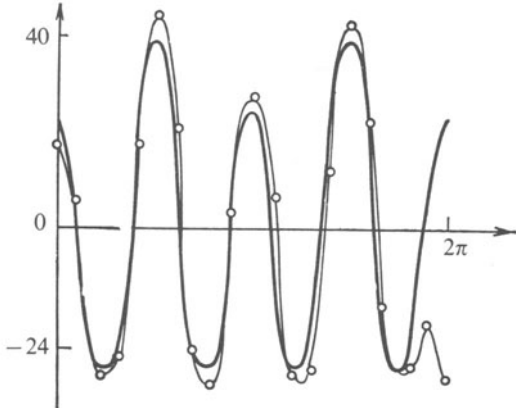

Figure 14

The first, second, and third derivatives of function $F(x)$ (bold line) and the corresponding spline approximations obtained using the method of structural risk minimization are presented in Figures 12, 13, and 14. These examples were solved using algorithm D-II.3 presented in Addendum II.

§7 The Problem of Probability Density Estimation

In Chapter 2 the problem of estimating a probability density in the class of continuous functions on $[a, b]$ was associated with the solution of an ill-posed problem of numerical differentiation. By definition a probability

density $f(t)$ is the derivative of a distribution function $F(x) = P(t \leq x)$, i.e., is a solution of the equation

$$\int_a^b \theta(x - t) f(t)\, dt = F(x). \tag{9.70}$$

Therefore the problem of estimating a density $f(t)$ on the basis of empirical data $t_1, \ldots, t_l$ should be viewed as one of an approximate solution to the integral equation (9.70), where the right-hand side is defined only approximately: instead of the distribution function $F(x)$, its estimator

$$F_l(x) = \frac{1}{l} \sum_{i=1}^{l} \theta(x - t_i)$$

is available.

We shall solve this problem using the regularization method (cf. Appendix to Chapter 1). We write the functional

$$R_{\gamma_l}(\hat{f}, F_l) = \rho^2_{E_2}(A\hat{f}, F_l) + \gamma_l \Omega(\hat{f}), \tag{9.71}$$

where A is the operator of Equation (9.70) and γ_l is a constant of regularization; $\gamma_l \to 0$ as $l \to \infty$.

Consider the sequence of elements

$$f_1^{\gamma_1}(t), \ldots, f_l^{\gamma_l}(t), \ldots \tag{9.72}$$

minimizing (9.71) as $l \to \infty$. This sequence is random, since it is formed using the random functions $F_l(x)$.

In the Appendix to this chapter a theorem is proved which asserts that if the desired solution of the operator equation belongs to a compact set $\Omega(f) \leq c$, then for any ν and μ there exists $n(\mu, \nu)$ such that the inequality

$$P\{\rho_{E_1}(f_l^{\gamma_l}, f) > \nu\} \leq P\{\rho^2_{E_2}(F_l, F) > \mu\gamma_l\} \tag{9.73}$$

is fulfilled for all elements of (9.72) starting with some $l > n(\mu, \nu)$. We shall utilize this inequality for defining conditions which will assure convergence of the sequence (9.72) to the desired density.

Consider the asymptotic bound on the rate of convergence of the empirical distribution function to the population distribution function (the Kolmogorov–Smirnov bound)

$$P\left\{\sup_x |F_l(x) - F(x)| > \varepsilon\right\} < 2e^{-2\varepsilon^2 l}. \tag{9.74}$$

Now let $\rho_{E_2}(F_l, F) = \sup_x |F_l(x) - F(x)|$. From (9.73) and (9.74) we obtain

$$P\{\rho_{E_1}(f_l^{\gamma_l}, f) > \nu\} < 2e^{-\mu l \gamma_l}.$$

It follows from this inequality that in order for the sequence (9.72) to converge in probability in the metric of the space E_1 to the population

density it is sufficient that

$$\begin{aligned} &\gamma_l \xrightarrow[l\to\infty]{} 0, \\ &l\gamma_l \xrightarrow[l\to\infty]{} \infty, \end{aligned} \tag{9.75}$$

and in order for the sequence to converge with probability 1 it is sufficient in view of the Borel–Cantelli lemma that at least for one μ the inequality

$$\sum_{l=1}^{\infty} e^{-\mu l \gamma_l} < \infty \tag{9.76}$$

be fulfilled. Using in (9.71) different stabilizing functionals $\Omega(f)$, one may obtain estimates $f_l^{\gamma_l}$ of the density which converge to the desired density in different metrics.

We have thus established that if a density belongs to a compactum $\Omega(f) \le c$, one can select γ_l such that the sequence (9.72) will converge to the desired density.

The requirement that the desired density belong to a compactum may be avoided. It is shown in the Appendix to this chapter that one can obtain a sequence of solutions converging to a continuous density—it is sufficient to choose as the stabilizing functional the functional $\Omega(f) = \|\hat{f}\|^2$, where $\|\hat{f}\|$ is the norm of a Hilbert space. But now the condition (9.75) is fulfilled for any positive μ and the sequence converges to the solution in metric L_2.

Thus the methods of density estimation are associated with solutions of ill-posed problems of numerical differentiation.

§8 Estimation of Smooth Densities

We shall apply the regularization method to estimate smooth densities defined on the interval $[a, b]$.

Suppose it is known that the probability density $f(t)$ possesses m derivatives (m may be equal to 0), and let the function $f^{(m)}(t)$ satisfy the Lipschitz condition of order μ $(0 < \mu \le 1)$:

$$|f^{(m)}(t) - f^{(m)}(\tau)| < K(f)|t - \tau|^{\mu},$$

$$K(f) = \sup_{t,\tau\in[a,b]} \frac{|f^{(m)}(t) - f^{(m)}(\tau)|}{|t - \tau|^{\mu}}.$$

Consider the following functional:

$$\Omega^*(\hat{f}) = \left(\max_{0\le k\le m} \sup_{t\in[a,b]} |\hat{f}^{(k)}(t)| + \sup_{t,\tau\in[a,b]} \frac{|\hat{f}^{(m)}(t) - \hat{f}^{(m)}(\tau)|}{|t - \tau|^{\mu}} \right)^2. \tag{9.77}$$

This functional is lower semicontinuous, and the set of functions

$$\mathcal{M}_c = \{\hat{f}: \Omega^*(\hat{f}) \le c\}$$

is a compact set in C. Therefore the functional (9.77) may be used for constructing the regularizing functional (9.71).

We choose for γ_l the sequence

$$\gamma_l = \frac{\ln \ln l}{l}.$$

This sequence satisfies the condition (9.75), and thus in view of the results of the preceding section the sequence of elements f_l which minimize the functional

$$R_l^*(\hat{f}; F_l) = \left(\sup_{x \in [a,b]} \left| \int_a^b \theta(x-t)\hat{f}(t)\,dt - F_l(x) \right| \right)^2 + \frac{\ln \ln l}{l} \Omega^*(\hat{f}) \quad (9.78)$$

converges in probability as l increases to the required density in the metric C.

In this section we shall estimate the asymptotic rate of convergence of the sequence of solutions f_l to the required density. As will be shown below, the rate of convergence depends on the degree of smoothness of the estimated density characterized by the quantity

$$\beta = m + \mu$$

(the larger the β the larger the rate).

Theorem 9.5 (Stefanjuk). *An asymptotic rate of convergence in the metric C of estimators of the density $f_l(t)$ to the required function $f(t)$ is determined by the expression*

$$P\left\{ \overline{\lim_{l \to \infty}} \left(\frac{l}{\ln \ln l} \right)^{\beta/2(\beta+1)} \sup_{t \in [a,b]} |f_l(t) - f(t)| \le g \right\} = 1,$$

where g is a constant.

Observe that in spite of the fact that the sequence (9.78) does not satisfy the conditions (9.76) obtained in the preceding section, Theorem 9.5 assures uniform convergence of $f_l(t)$ to $f(t)$ with probability 1. The condition (7.6) is only a sufficient condition. The result of Theorem 9.5 is more refined.

Finally, the following should be mentioned before proceeding to the proof of the theorem.

R. Z. Has'minskii obtained in 1978 an estimate for the best rate of convergence of an approximation to an arbitrary density [113]. He discovered there is no algorithm which would ensure convergence in $C[a, b]$ to a β-smooth density at a rate whose order of magnitude is larger than $(l/\ln l)^{-\beta/2\beta+1}$.

As soon as this result became available, an attempt was made to improve on Theorem 9.5 (namely to obtain the order of magnitude of the rate $(l/\ln l)^{-\beta/2\beta+1}$ instead of $(l/\ln l)^{-\beta/2\beta+2}$ as given in Theorem 9.5.) However, for approximations that are generated by functional $R_l^*(f; F_l)$ such attempts have failed.

The best rate of convergence in terms of the order of magnitude was obtained for another sequence generated by minimizing a modified functional. Such a functional is constructed as follows: Subdivide the segment $[a, b]$ into $n = [(l/\ln l)^{1/2\beta+1}]$ equal parts

$$[x_i, x_{i-1}], x_i = a + i\frac{b-a}{n}, \qquad i = 1, \ldots, n,$$

and define the quantities:

$$\|\Phi(x)\|_i = |\Phi(x_i) - \Phi(x_{i-1})|.$$

Using these quantities the functional is then given by:

$$\hat{R}_l^*(f; F_l) = \left(\sup_{1 \le i \le n} \left\| \int_a^b \theta(x - t) f(t)\, dt - F_l(x) \right\|_i \right)^2 + \frac{\ln l}{l} \Omega^*(f).$$

There is a theorem (its proof is analogous to that of Theorem 9.5) which asserts that a sequence of elements $\hat{f}_l(t)$ minimizing $\hat{R}_l^*(f; F_l)$ converges as l increases, in $C[a, b]$, to a β-smooth density $f(t)$ at a rate whose order of magnitude is the best obtainable:

$$P\left\{ \overline{\lim_{l \to \infty}} \left(\frac{l}{\ln l} \right)^{\beta/(2\beta+1)} \sup_{a \le t \le b} |\hat{f}_l(t) - f(t)| \le g \right\} = 1.$$

However the functional $\hat{R}_l^*(f; F_l)$ does not satisfy the requirements of the general theory of solving ill-posed problems using the method of regularization. It is constructed by means of the value of $\sup_i \|\Phi(x)\|_i$ which does not specify a norm in $C[a, b]$.

It should also be noted that the best possible rate of convergence may also be obtained for other methods, e.g., for Parzen's method (see Section 9). However, for the latter method, the maximal rate was achieved for special constructions (such as Dirichlet kernels) only rather than for constructions (kernels) which are easy to handle and are usually employed.

These are the "racing" aspects of the problem.

To prove the theorem the following lemma will be required.

Lemma. *Consider a function $y(t)$ continuously differentiable on the interval $[a, b]$. Denote by $x(t)$ the derivative of this function. Let the mth $(m \ge 0)$ derivative of $x(t)$ satisfy the Lipschitz condition of order μ on $[a, b]$:*

$$\sup_{t, \tau \in [a, b]} |x^{(m)}(t) - x^{(m)}(\tau)| \le K |t - \tau|^{\mu}.$$

Then the inequality

$$\|x\|_C \le \max\{C_m^* \|y\|_C; C_m^{**} \|y\|_C^{(m+\mu)/(1+m+\mu)}\}$$

is valid, where

$$\|x\|_C = \sup_{t \in [a, b]} |x(t)|,$$

$$C_m^* = \frac{2^{1+2m}}{b-a} \left(\frac{1+\mu}{\mu} \right), \qquad C_m^{**} = \left[2^{(m+\mu)(1+\mu+m)-\mu^2} K \cdot \left(\frac{\mu+1}{\mu} \right)^{\mu} \right]^{1/(1+m+\mu)}.$$

PROOF.

(1) Consider first the case of $m = 0$. Choose on $[a, b]$ an arbitrary point t^* such that $|x(t^*)| \neq 0$. Define an ε-neighborhood of this point with

$$\varepsilon = \left(\frac{|x(t^*)|}{K}\right)^{1/\mu}. \tag{9.79}$$

Assume that at least one of the endpoints of this ε-neighborhood—say the right one—is located within the interval $[a, b]$, i.e., $t^* + \varepsilon \leq b$. Along with the function $x(t)$ consider the function

$$\varphi(\tau) = |x(t^*)| - K \cdot (\tau - t^*)^{\mu}.$$

Since for any $\tau \in [t^*, t^* + \varepsilon]$

$$|x(t^*)| - |x(\tau)| \leq |x(t^*) - x(\tau)| \leq K \cdot (\tau - t^*)^{\mu},$$

it follows that

$$|x(\tau)| \geq |x(t^*)| - K \cdot (\tau - t^*)^{\mu} = \varphi(\tau). \tag{9.80}$$

Noting that ε is defined by (9.79), we conclude from (9.80) that on the interval $[t^*, t^* + \varepsilon]$ the function $x(\tau)$ remains of the same sign. Therefore the relation

$$|y(t^* + \varepsilon) - y(t^*)| = \left| \int_{t^*}^{t^* + \varepsilon} x(\tau)\, d\tau \right| = \int_{t^*}^{t^* + \varepsilon} |x(\tau)|\, d\tau \geq \int_{t^*}^{t^* + \varepsilon} \varphi(\tau)\, d\tau$$

$$= |x(t^*)|\varepsilon - K\varepsilon^{1+\mu} \frac{1}{1+\mu} = (K)^{-1/\mu} \left(\frac{\mu}{1+\mu}\right) |x(t^*)|^{(1+\mu)/\mu}$$

is valid. Since, however, the inequality

$$|y(t^* + \varepsilon) - y(t^*)| \leq 2\|y\|_C$$

is always fulfilled, it follows from the bound obtained that

$$|x(t^*)| \leq \left[2\left(\frac{1+\mu}{\mu}\right) K^{1/\mu} \|y\|_C\right]^{\mu/(1+\mu)}. \tag{9.81}$$

Now let both endpoints of the above mentioned ε-neighborhood of the point t^* be located outside the interval $[a, b]$. Consider also the function

$$\varphi_1(\tau) = \begin{cases} |x(t^*)| - |x(t^*)| \left(\dfrac{t^* - \tau}{t^* - a}\right)^{\mu} & \text{for } a \leq \tau \leq t^*, \\ |x(t^*)| - |x(t^*)| \left(\dfrac{\tau - t^*}{b - t^*}\right)^{\mu} & \text{for } t^* < \tau \leq b. \end{cases}$$

It is easy to verify that for any $\tau \in [a, b]$ the inequality $0 \leq \varphi_1(\tau) \leq |x(\tau)|$ is fulfilled. Therefore as above we have

$$|y(b) - y(a)| = \left| \int_a^b x(t)\, dt \right| = \int_a^b |x(t)|\, dt \geq \int_a^b \varphi_1(\tau)\, d\tau = \frac{\mu}{1+\mu} (b - a)|x(t^*)|.$$

Hence

$$|x(t^*)| \leq \frac{2}{b-a} \left(\frac{1+\mu}{\mu}\right) \|y\|_C. \tag{9.82}$$

Thus if at least one of the endpoints of the ε-neighborhood is located within the interval $[a, b]$, the bound (9.81) is valid; otherwise (9.82) is. While the inequalities were obtained for any t^* such that $|x(t^*)| \neq 0$, the bound

$$\|x\|_C \leq \max\{C_0^* \|y\|_C;\ C_0^{**} \|y\|_C^{\mu/(1+\mu)}\} \tag{9.83}$$

holds, where

$$C_0^* = \frac{2}{b-a}\left(\frac{1+\mu}{\mu}\right), \qquad C_0^{**} = \left[2\left(\frac{\mu+1}{\mu}\right)K^{1/\mu}\right]^{\mu/(1+\mu)}.$$

For the case $m = 0$ the lemma is thus proved.

(2) Now consider the bound

$$\|x^{(m-i)}\|_C \leq \max\{C_i^* \|x^{(m-i-1)}\|_C;\ C_i^{**} \|x^{(m-i-1)}\|_C^{(1+\mu)/(1+i+\mu)}\}. \tag{9.84}$$

This bound was obtained above for the case $i = 0$. For the case $i = m$ it constitutes the assertion of the lemma (here we use the notation $x^{(-1)}(t) = y(t)$). We shall prove the validity of (9.84) for $i = 1, 2, \ldots, m$ by induction.

Let the bound (9.84) be valid for $i = k - 1$. We show that it remains valid for $i = k$ as well. Indeed, since $x^{(m-k)}(t)$ is differentiable on $[a, b]$, we have

$$\sup_{t,\tau\in[a,b]} |x^{(m-k)}(t) - x^{(m-k)}(\tau)| \leq \|x^{(m-k+1)}\|_C |t - \tau|;$$

hence the function $x^{(m-k)}(t)$ satisfies the Lipschitz condition of order $\mu = 1$. Therefore utilizing (9.83) we obtain

$$\|x^{(m-k)}\|_C \leq \max\left\{\frac{2^2}{b-a}\|x^{(m-k-1)}\|_C;\ 2\|x^{(m-k+1)}\|_C^{1/2}\|x^{(m-k-1)}\|_C^{1/2}\right\}.$$

By the induction assumption

$$\|x^{(m-k+1)}\|_C \leq \max\{C_{k-1}^* \|x^{(m-k)}\|_C;\ C_{k-1}^{**} \|x^{(m-k)}\|_C^{(k-1+\mu)/(k+\mu)}\}$$

Combining these two inequalities, we have

$$\begin{aligned}\|x^{(m-k)}\|_C \leq \max\Big\{&\frac{2^2}{b-a}\|x^{(m-k-1)}\|_C;\\ &2[C_{k-1}^* \|x^{(m-k)}\|_C]^{1/2}\|x^{(m-k-1)}\|_C^{1/2};\\ &2[C_{k-1}^{**}\|x^{(m-k)}\|_C]^{\mu+k-1/2(\mu+k)}\|x^{(m-k-1)}\|_C^{1/2}\Big\}.\end{aligned} \tag{9.85}$$

It follows from (9.85) that

$$\begin{aligned}\|x^{(m-k)}\|_C \leq \max\Big\{&\frac{2^2}{b-a}\|x^{(m-k-1)}\|_C;\ \frac{2}{b-a}C_{k-1}^*\|x^{(m-k-1)}\|_C;\\ &(4C_{k-1}^{**})^{(\mu+k)/(\mu+k+1)}\|x^{(m-k-1)}\|_C^{(\mu+k)/(\mu+k+1)}\Big\}.\end{aligned}$$

Finally, taking the values of C_k^* and C_k^{**} into account, we arrive at the inequality

$$\|x^{(m-k)}\|_C \leq \max\{C_k^* \|x^{(m-k-1)}\|_C;\ C_k^{**}\|x^{(m-k-1)}\|_C^{(k+\mu)/(k+\mu+1)}\}.$$

For $k = m$ the inequality obtained is the assertion of the lemma. □

PROOF OF THE THEOREM. According to Smirnov's formula the deviation between the empirical distribution function $F_l(x)$ and the population distribution function satisfies with probability 1 the relation

$$\overline{\lim_{l\to\infty}}\left(\frac{2l}{\ln\ln l}\right)^{1/2}\|F_l(x)-F(x)\|_C=1.$$

Therefore for any ε there exists $N=N(\varepsilon)$ such that simultaneously for all $l>N$ the inequality

$$\frac{l}{\ln\ln l}\|F_l(x)-F(x)\|_C^2<1 \tag{9.86}$$

is fulfilled with probability $1-\varepsilon$.

Let $f_l(t)$ be the function which minimizes the functional (9.78), and let $f(t)$ be the desired density. Then

$$\frac{\ln\ln l}{l}\Omega^*(f_l)\le R_l^*(f_l,F_l)\le R_l^*(f,F_l)$$

$$=\left\|\int_a^b\theta(x-t)f(t)\,dt-F_l(x)\right\|_C^2+\frac{\ln\ln l}{l}\Omega^*(f),$$

whence we obtain

$$\Omega^*(f_l)\le\Omega^*(f)+\left\|\int_a^b\theta(x-t)f(t)\,dt-F_l(x)\right\|_C^2\frac{l}{\ln\ln l}.$$

Observe that starting with $l=N(\varepsilon)$, the inequality (9.86) is fulfilled with probability $1-\varepsilon$; hence starting with $N(\varepsilon)$, the inequality

$$\Omega^*(f_l)\le\Omega^*(f)+1 \tag{9.87}$$

is satisfied with probability $1-\varepsilon$. If the mth derivative of the desired density $f(t)$ satisfies the Lipschitz condition of order μ and the functional $\Omega^*(f)$ is (9.77), then it follows from (9.87) that

$$\sup_{t,\tau}\frac{|f_l^{(m)}(t)-f_l^{(m)}(\tau)|}{|t-\tau|^\mu}\le(\Omega^*(f)+1)^{1/2},$$

i.e., the mth derivative of the function $f_l(t)$ satisfies with probability $1-\varepsilon$ the Lipschitz condition of order μ with the constant $K=(\Omega^*(f)+1)^{1/2}$. Therefore in view of the lemma, the inequality

$$\|f_l-f\|_C\le\max\left\{C_m^*\left\|\int_a^b\theta(x-t)f_l(t)\,dt-F(x)\right\|_C;\right.$$

$$\left.C_m^{**}\left\|\int_a^b\theta(x-t)f_l(t)\,dt-F(x)\right\|_C^{\beta/(1+\beta)}\right\}\quad(\beta=m+\mu) \tag{9.88}$$

is valid with probability $1-\varepsilon$. Multiplying both sides of the inequality by

$$\left(\frac{l}{\ln\ln l}\right)^{\beta/2(1+\beta)},$$

we obtain

$$\left(\frac{l}{\ln\ln l}\right)^{\beta/2(1+\beta)} \|f_l - f\|_C \le \max\left\{C_m^*\left(\sqrt{\frac{\ln\ln l}{l}}\right)^{1/(1+\beta)} \times \left[\sqrt{\frac{l}{\ln\ln l}}\left\|\int_a^b \theta(x-t)f_l(t)\,dt - F(x)\right\|_C\right]; C_m^{**}\left[\sqrt{\frac{l}{\ln\ln l}}\left\|\int_a^b \theta(x-t)f_l(t)\,dt - F(x)\right\|_C\right]^{\beta/(1+\beta)}\right\}. \tag{9.89}$$

Observe now that starting with $N(\varepsilon)$ the inequality

$$\sqrt{\frac{l}{\ln\ln l}}\left\|\int_a^b \theta(x-t)f_l(t)\,dt - F(x)\right\|_C \le 1 + \sqrt{\Omega^*(f)+1} \tag{9.90}$$

is fulfilled for all l with probability $1-\varepsilon$. The inequality (9.90) follows from the triangle inequality

$$\left\|\int_a^b \theta(x-t)f_l(t)\,dt - F(x)\right\|_C \le \left\|\int_a^b \theta(x-t)f_l(t)\,dt - F_l(x)\right\|_C + \|F_l(x) - F(x)\|_C,$$

the self-evident system of inequalities

$$\left\|\int_a^b \theta(x-t)f_l(t)\,dt - F_l(x)\right\|_C^2 \le R_l^*(f_l, F_l) \le R_l^*(f, F_l),$$

and the bound (9.86).

Taking (9.89) and (9.90) into account, we may assert that with probability $1-\varepsilon$ for all $l \ge N(\varepsilon)$ the inequality

$$\left(\frac{l}{\ln\ln l}\right)^{\beta/2(1+\beta)} \|f_l - f\|_C \le \max\left\{C_m^{***}\left(\sqrt{\frac{\ln\ln l}{l}}\right)^{1/(1+\beta)}; g\right\} \tag{9.91}$$

is fulfilled, where

$$C_m^{***} = C_m^*(1 + \sqrt{\Omega^*(f)+1}),$$
$$g = C_m^{**}(1 + \sqrt{\Omega^*(f)+1})^{\beta/(1+\beta)}.$$

Evidently, starting with some number N^*, the inequality

$$C_m^{***}\left(\sqrt{\frac{\ln\ln l}{l}}\right)^{1/(1+\beta)} < g$$

will be satisfied. Thus starting with $N^*(\varepsilon) = \max(N^*, N(\varepsilon))$, with probability $1-\varepsilon$ the inequality

$$\left(\frac{l}{\ln\ln l}\right)^{\beta/2(1+\beta)} \|f_l - f\|_C < g \tag{9.92}$$

will be fulfilled. Since for any ε there exists $N^*(\varepsilon)$ such that for all $l \ge N^*(\varepsilon)$ simultaneously the inequality (9.92) is fulfilled with probability $1-\varepsilon$, we have with probability 1

$$\overline{\lim_{l\to\infty}}\left(\frac{l}{\ln\ln l}\right)^{\beta/2(1+\beta)} \|f_l - f\|_C < g.$$

The theorem is proved. □

Below we shall also utilize the method of structural minimization of the risk for solving problems of density estimation. However before proceeding to the exposition of the corresponding results it should be noted that there exist nonparametric methods of estimating density (for example Parsen's method) which seem to avoid solving an ill-posed problem. However, a more detailed analysis of these methods shows that all of them involve a constant whose determination is a problem which is completely equivalent to the determination of the constant of regularization γ_l when solving ill-posed problems.

§9 Density Estimation Using Parzen's Method

Parzen's idea for estimating a density is as follows: The identity

$$P(x) = \int \delta(x - t)P(t)\, dt$$

is valid. Consider a parametric sequence of functions converging to $\delta(x)$:

$$\frac{1}{h_1} K\left(\frac{x}{h_1}\right), \ldots, \frac{1}{h_l} K\left(\frac{x}{h_l}\right);$$

$$\lim_{l \to \infty} \frac{1}{h_l} K\left(\frac{x}{h_l}\right) = \delta(x).$$

Such a sequence does exist. For example,

$$\lim_{h \to 0} \frac{1}{\sqrt{2\pi} h} e^{-x^2/2h^2} = \delta(x).$$

For any continuous density $P(x)$ there exists a quantity h such that replacing $\delta(x)$ by $(1/n)K(x/n)$ in the integrand hardly affects the result, i.e.,

$$P(x) = \int \delta(x - t)P(t)\, dt \approx \int \frac{1}{h} K\left(\frac{x - t}{h}\right) P(t)\, dt.$$

We now replace the mathematical expectation by the value of the sample mean. For a sufficiently large sample size this will affect the result only slightly:

$$P(x) \approx \int \frac{1}{h} K\left(\frac{x - t}{h}\right) P(t)\, dt \approx \frac{1}{l} \sum_{i=1}^{l} \frac{1}{h} K\left(\frac{x - t_i}{h}\right).$$

The expression on the right-hand side is thus used as the formula for density estimation:

$$\hat{P}_l(x) = \frac{1}{l} \sum_{i=1}^{l} \frac{1}{h} K\left(\frac{x - t_i}{h}\right). \tag{9.93}$$

The problem is thus:

(1) What should the law be for forming the quantity h so that as the sample size increases the estimator tends to the true probability density?
(2) How should the constant h be chosen if the sample size is finite?

There is no answer to the second question. As far as the asymptotic properties of this method are concerned, Parzen in 1962 and Nadaraya in 1965 obtained conditions which assure the convergence of the estimator (9.93) to the desired uniformly continuous density. It turns out that for convergence in metric C of the sequence (9.93) in probability to the desired density it is sufficient that

$$\begin{aligned} h_l &\xrightarrow[l\to\infty]{} 0, \\ lh_l^2 &\xrightarrow[l\to\infty]{} \infty \end{aligned} \tag{9.94}$$

(Parzen's result), while for convergence with probability 1 it is sufficient that for any positive μ the series

$$\sum_{l=1}^{\infty} e^{-\mu h_l^2 l} < \infty \tag{9.95}$$

converge (Nadaraya's result).

Observe that the conditions (9.94) and (9.95) for choosing constants h_l are equivalent to the conditions (9.75) and (9.76) for choosing regularizing constants when solving the ill-posed problem (9.70). This basically means that when circumventing an ill-posed problem it is impossible to avoid the difficulties connected with its solution.

Parzen's method may be obtained directly as a solution of Equation (9.70) using the regularization method.

Let the desired density $f(t)$ be square-integrable on $(-\infty, +\infty)$. In accordance with the regularization method we shall obtain a function $f_l(t) \in L_2$ which minimizes the functional

$$R_l^*(\hat{f}, F_l) = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} \theta(x-t)\hat{f}(t)\,dt - F_l(x) \right]^2 dx + \gamma_l \int_{-\infty}^{+\infty} \hat{f}^2(t)\,dt.$$

From the minimum condition for functional $R_l^*(\hat{f}, F_l)$ at the point $f_l(t)$ we obtain

$$\int_{-\infty}^{+\infty} \theta(x-t) \left[\int_{-\infty}^{+\infty} \theta(x-\tau) f_l(\tau)\,d\tau - F_l(x) \right] dx + \gamma_l f_l(t) = 0.$$

We solve this equation using the method of the Fourier transform for generalized functions. To do this we apply the Fourier transform to the equality, taking into account that

$$F_l(x) = \frac{1}{l} \sum_{k=1}^{l} \theta(x - t_k)$$

and the Fourier transform for the generalized function $\theta(x)$ is

$$\frac{i}{\omega} + \pi\delta(\omega).$$

We obtain the equation

$$\left(\frac{1}{i\omega}\right)\left[\left(-\frac{1}{i\omega}\right)\Pi(\omega) - \frac{1}{l}\sum_{k=1}^{l}\left(-\frac{e^{i\omega t_k}}{i\omega}\right)\right] + \gamma_l \Pi(\omega) = 0,$$

in which

$$\Pi(\omega) = \int_{-\infty}^{+\infty} f_l(t) e^{-i\omega t}\, dt$$

is the Fourier transform of the function $f_l(t)$. Next it follows that

$$\Pi(\omega) = \frac{\frac{1}{l}\sum_{k=1}^{l} e^{i\omega t_k}}{1 + \gamma_l \omega^2}.$$

Applying the inverse Fourier transform to $\Pi(\omega)$, we arrive at the estimate of the density

$$f_l(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Pi(\omega) e^{i\omega t}\, d\omega = \frac{1}{l}\sum_{k=1}^{l} \frac{1}{2\sqrt{\gamma_l}} e^{-|t-t_k|/\sqrt{\gamma_l}}.$$

We have thus obtained Parzen's estimator with the kernel

$$K(t) = \frac{1}{2\sqrt{\gamma_l}} \exp\left\{-\frac{|t|}{\sqrt{\gamma_l}}\right\}.$$

By using other stabilizing functionals $\Omega(f)$ one can obtain other kernels $K(t)$ as well.

§10 Density Estimation Using the Method of Structural Risk Minimization

We shall solve Equation (9.70) where in place of $F(x)$ the empirical estimator $F_l(x)$ is utilized. Observe that $F_l(x)$ is a random function whose values at two distinct points x_i^* and x_j^* are correlated.

The covariance coefficient K_{ij} of the random variables $y_i = F_i(x_i^*)$ and $y_j = F_l(x_j^*)$ equals

$$K_{ij} = \begin{cases} \frac{1}{l} F_i(1 - F_j) & \text{if } x_i^* \le x_j^*, \\ \frac{1}{l} F_j(1 - F_i) & \text{if } x_j^* < x_i^*. \end{cases} \tag{9.96}$$

Here we denote $F_i = F(x_i^*)$ and $F_j = F(x_j^*)$.

Consider N variables $y_1, \ldots, y_N$ formed by means of the function $F_l(x)$, and N random numbers $x_1^*, \ldots, x_N^*$ generated by a probability density uniform on $[a, b]$: $y_i = F_l(x_i^*)$, $i = 1, 2, \ldots, N$.

Since the random variables y_i and y_j are correlated, one cannot apply directly the method of structural minimization for estimating the probability

density. Therefore we apply to the random vector $Y = (y_1, \ldots, y_N)^T$ the following linear transformation:

$$Z = BY, \qquad B^T B = K^{-1},$$

where K is a covariance matrix with elements (9.96). It is known that by means of this transformation one can form a random vector z whose components are uncorrelated and have a unit variance. Therefore one can consider each component z_i of the vector z as a realization of a random variable conditioned on x_i^* in an independent sample of size N.

Denote by F_α a vector with coordinates $F(x_1^*, \alpha), \ldots, F(x_N^*, \alpha)$. The transformation B maps the vector F_α into the vector $W = BF_\alpha$, whose components are considered as values of a function $W(x, \alpha)$ at points $x_1^*, \ldots, x_N^*$.

We shall now solve the problem of minimizing the functional

$$\hat{I}(\alpha) = \int (z - W(x, \alpha))^2 P(z|x)\, dz\, dx.$$

To solve this problem we shall use the method of structural risk minimization. Thus we have

$$\hat{I}(\alpha) < \left[\frac{\frac{1}{N} \sum_{i=1}^{N} (z_i - W(x_i^*, \alpha))^2}{1 - 2\tau a(p) \sqrt{\frac{\ln m^s(2N) - \ln(\eta/8)}{N}}} \right]_\infty .$$

The numerator on the right-hand side can be represented as

$$\frac{1}{N} \sum_{i=1}^{N} (z_i - W(x_i^*, \alpha))^2 = \frac{1}{N} (Y - F_\alpha)^T B^T B (Y - F_\alpha)$$

$$= \frac{1}{N} (Y - F_\alpha)^T K^{-1} (Y - F_\alpha).$$

Finally we write the functional whose minimization over the classes of functions $S_1 \subset \cdots \subset S_N$ and over all the functions $F(x, \alpha)$ in each one of the classes determines an estimate of the probability density function

$$R(\alpha) = \left[\frac{\frac{1}{N} (Y - F_\alpha)^T K^{-1} (Y - F_\alpha)}{1 - 2\tau a(p) \sqrt{\frac{\ln m^{S_p}(2N) - \ln(\eta/8)}{N}}} \right]_\infty , \tag{9.97}$$

where

$$F_\alpha = (F(x_1^*, \alpha), \ldots, F(x_N^*, \alpha))^T,$$

$$F(x_i^*, \alpha) = \int_{-\infty}^{x_i^*} f(t, \alpha)\, dt,$$

$f(t, \alpha)$ being functions belonging to S_p.

In order to utilize (9.97) it is necessary to know the inverse of the covariance matrix. This matrix may be obtained analytically.† A direct verification of relation $KK^{-1} = \mathrm{I}$ yields that the following matrix is the inverse for K:

$$K^{-1} = \begin{Vmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \cdot & \cdot & \cdots & \cdot \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{Vmatrix}, \tag{9.98}$$

where

$$a_{ij} = 0 \quad \text{for } |i-j| \geq 2,$$

$$a_{i\,i-1} = a_{i-1\,i} = \frac{-l}{F_i - F_{i-1}},$$

$$a_{ii} = \frac{l(F_{i+1} - F_{i-1})}{(F_{i+1} - F_i)(F_i - F_{i-1})}.$$

However, the matrix (9.98) is expressed in terms of the unknown distribution function of the random variable $F(x)$ whose derivative we are to obtain. It thus turns out that in order to estimate a probability density based on a sample it is necessary to possess some prior information about the density (i.e., to know the covariance matrix (9.98)). This is the fundamental difficulty of the problem of estimating a probability density in a wide class of functions.

In place of matrix (9.98) one can utilize in (9.97) its estimator where the values of the matrix are determined by means of a preliminary estimated continuous function $F_{\text{emp}}(x)$. Observe that the problem of estimating the function $F(x)$ is simpler than the problem of density estimation (in view of the Glivenko–Cantelli theorem, the empirical distribution function converges to the population one in the uniform metric). Thus the algorithm of density estimation consists of two stages: the preliminary estimation of the matrix (9.98) and the determination of the density.

We estimate the matrix using the polygon $F_{\text{emp}}(x)$ of the distribution function‡ constructed from the order statistics $x_1^*, \ldots, x_l^*$ generated by the sample $t_1, \ldots, t_l$:

$$F_{\text{emp}}(x) = \begin{cases} \dfrac{1}{2l}\dfrac{x}{x_1^* - a} & \text{if } a < x \leq x_1^*, \\ \dfrac{k - \frac{1}{2}}{l} + \dfrac{1}{2l}\dfrac{x - x_k^*}{x_{k+1}^* - x_k^*} & \text{if } x_k^* < x \leq x_{k+1}^*,\ k = 1, \ldots, l-1, \\ \dfrac{l - \frac{1}{2}}{l} + \dfrac{1}{2l}\dfrac{x - x_l^*}{b - x_l^*} & \text{if } x_l^* < x \leq b. \end{cases}$$

† We shall assume that the required density on (a, b) does not vanish.

‡ One can determine the relation between the sample size l and the number of points N at which the value of the polygon $F_{\text{emp}}(x)$ is determined such that in the two-stage method of estimation the function approaches the desired one as the sample size increases.

In Figures 15, 16, and 17 are presented the results of estimating various densities using the Parzen method (a) and the two-stage method described above (b). The actual densities are the bold lines; the approximations obtained are the dotted lines. Figure 15 shows the estimation of a unimodal density based on a sample of size $l = 50$. Figure 16 shows the estimation of a bimodal density based on a sample of size $l = 50$. Finally, the estimation of a trimodal density based on a sample of size $l = 100$ is presented in Figure 17.

In the case of the Parzen estimator the kernel $(1/\sqrt{2\pi}h)e^{-(t-\tau)^2/2h^2}$ was used. The value of h was chosen to be equal to $h = \hat{\sigma}l^{-1/5}$, where $\hat{\sigma}^2$ is the sample variance (this is the usually recommended value of h).

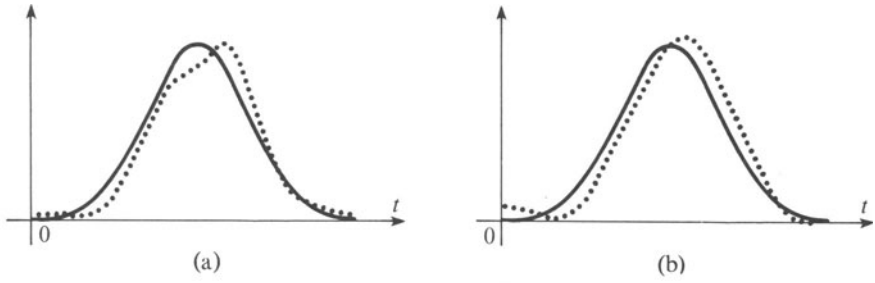

Figure 15

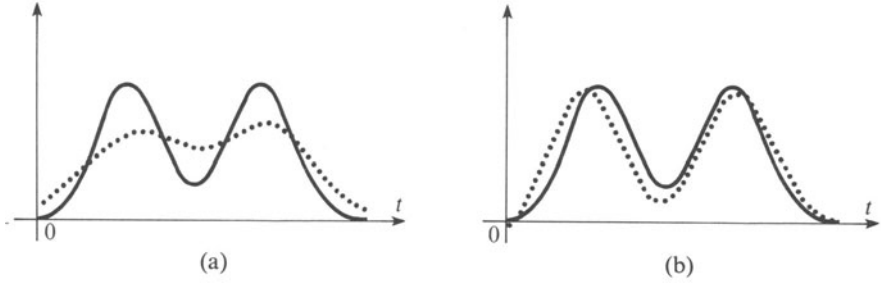

Figure 16

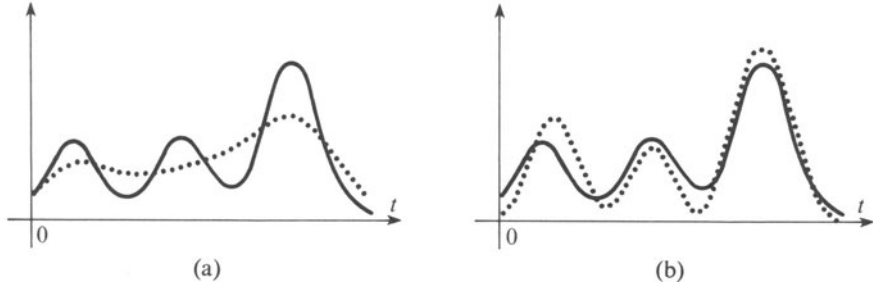

Figure 17

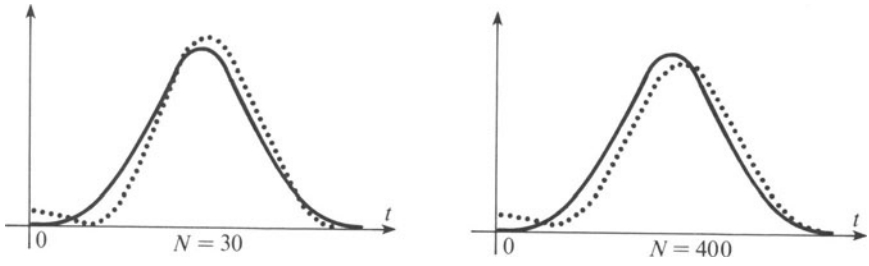

Figure 18

In the case of estimation by means of the two-stage procedure the value of the polygon $F_{\text{emp}}(x)$ at $N = l$ randomly selected points was used. Variation of N even over wide limits affects the result only slightly. Estimation of the density based on 50 points for $N = 30$ and $N = 400$ is presented in Figure 18. The densities were estimated using the modified algorithm D-II.3 (see Addendum 2).

As can be seen from these examples, Parzen's method may lead, in the case of finite sample sizes l, to unsatisfactory results if the desired density is multimodal.† Under the same conditions the two-stage method allows us to obtain sufficiently good estimators for unimodal as well as for multimodal probability densities.

† In the case of estimating a trimodal density using Parzen's method the same accuracy (closeness in metric L_2) as the one shown in Figure 17(b) can be achieved only if the sample size is increased more than tenfold (for $l = 1200$).

Appendix to Chapter 9

Statistical Theory of Regularization

Consider the operator equation

$$Af = F \tag{A.1}$$

defined by a continuous operator A which maps in a one-to-one manner the elements of a metric space E_1 into elements of the metric space E_2. Let $\Omega(f)$ be a lower semicontinuous functional such that:

(1) a solution of Equation (A.1) belongs to $D(\Omega)$, the domain of definition of the functional $\Omega(f)$;
(2) the functional $\Omega(f)$ takes on real nonnegative values in $D(\Omega)$;
(3) the sets $\mathcal{M}_c = \{f: \Omega(f) \leq c\}$, $c > 0$, are compact.

Consider random functions F_l and elements $f_l^{\gamma_l}$ which minimize the functional

$$R_{\gamma_l}(\hat{f}, F_l) = \rho_{E_2}^2(A\hat{f}, F_l) + \gamma_l \Omega(\hat{f}). \tag{A.2}$$

Let $\gamma_l \to 0$ as $l \to \infty$. Under these conditions the following two theorems, which are stochastic analogs of A. N. Tihonov's theorems, are valid (see the Appendix to Chapter 1, Theorems A.1 and A.2).

Theorem A.1. *For any positive numbers ν and μ there exists a positive number $n(\mu, \nu)$ such that for all $l > n(\mu, \nu)$ the inequalities*

$$P\{\rho_{E_1}(f_l^{\gamma_l}, f) > \nu\} \leq P\{\rho_{E_2}^2(F_l, F) > \mu\gamma_l\}$$

are fulfilled.

Theorem A.2. *Let E_1 be a Hilbert space, A be a linear operator, and $\Omega(\hat{f}) = \| \hat{f} \|^2$; then for any ε there exists a number $n(\varepsilon)$ such that for all $l > n(\varepsilon)$ the*

inequalities

$$P\{\|f_l^{\gamma_l} - f\|^2 > \varepsilon\} \leq 2P\left\{\rho_{E_2}^2(F_l, F) > \frac{\varepsilon}{2}\gamma_l\right\}$$

are fulfilled.

PROOF OF THEOREM A.1. By definition, for any l the chain of inequalities†

$$\begin{aligned}\gamma_l\Omega(f_l^{\gamma_l}) &\leq R_{\gamma_l}(f_l^{\gamma_l}, F_l) \leq R_{\gamma_l}(f, F_l)\\ &= \rho_2^2(Af, F_l) + \gamma_l\Omega(f) = \rho_2^2(F, F_l) + \gamma_l\Omega(f)\end{aligned} \tag{A.3}$$

is valid. In other words, the inequality

$$\Omega(f_l^{\gamma_l}) \leq \Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \tag{A.3a}$$

is valid. Moreover, clearly

$$\rho_2^2(Af_l^{\gamma_l}, F_l) \leq R_{\gamma_l}(f_l^{\gamma_l}, F_l). \tag{A.4}$$

Utilizing (A.3) and (A.4), we obtain the inequalities

$$\begin{aligned}\rho_2(Af_l^{\gamma_l}, F) &\leq \rho_2(Af_l^{\gamma_l}, F_l) + \rho_2(F_l, F)\\ &\leq \rho_2(F_l, F) + \sqrt{\rho_2^2(F_l, F) + \gamma_l\Omega(f)}.\end{aligned} \tag{A.5}$$

Furthermore for any $v > 0$ and $c > \Omega(f)$ the equality

$$\begin{aligned}P\{\rho_1(f_l^{\gamma_l}, f) \leq v\} &= P\left\{\rho_1(f_l^{\gamma_l}, f) \leq v \mid \Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leq c\right\}\\ &\quad \times P\left\{\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leq c\right\}\\ &\quad + P\left\{\rho_1(f_l^{\gamma_l}, f) \leq v \mid \Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} > c\right\}\\ &\quad \times P\left\{\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} > c\right\}\end{aligned} \tag{A.6}$$

is valid. Now let the condition

$$\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leq c$$

be fulfilled. Then it follows from (A.3a) that the inequality $\Omega(f_l^{\gamma_l}) \leq c$ is valid, i.e., $f_l^{\gamma_l}$ belongs to a compactum. In view of the lemma on the continuity of the inverse operator for A on a compactum (Appendix to Chapter 1), we obtain that there exists a δ such that the inequality

$$\rho_1(f_l^{\gamma_l}, f) \leq v$$

† Here and below we set $\rho_{E_i} = \rho_i$ for notational simplicity.

is fulfilled as long as the inequality $\rho_2(Af_l^{\gamma_l}, F) \leq \delta$ is. Hence we have for l sufficiently large that

$$P\left\{\rho_1(f_l^{\gamma_l}, f) \leq v|\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leq c\right\}$$

$$\geq P\left\{\rho_2(Af_l^{\gamma_l}, F_l) \leq \delta|\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leq c\right\}. \quad \text{(A.7)}$$

Observe now that in view of (A.5) the inequality

$$\rho_2(Af_l^{\gamma_l}, F) \leq \sqrt{\gamma_l(c - \Omega(f))} + \sqrt{\gamma_l(c - \Omega(f)) + \gamma_l\Omega(f)}$$
$$= \sqrt{\gamma_l}(\sqrt{c - \Omega(f)} + \sqrt{c})$$

is fulfilled in the domain

$$\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leq c.$$

Since $\gamma_l \to 0$ as $l \to \infty$ for any δ starting with some n, the inequality

$$P\left\{\rho_2(Af_l^{\gamma_l}, F) \leq \delta|\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leq c\right\} = 1$$

is fulfilled for all $l > n$. And since (A.7) is valid for all $l > n$, the equality

$$P\left\{\rho_1(f_l^{\gamma_l}, f) \leq v|\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leq c\right\} = 1$$

is fulfilled. Thus it follows from (A.6) that for any $v > 0$ there exists n such that for $l > n$ the inequality

$$P\{\rho_1(f_l^{\gamma_l}, f) \leq v\} > P\left\{\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leq c\right\}$$

is fulfilled, and hence also the inequality

$$P\{\rho_1(f_l^{\gamma_l}, f) > v\} \leq P\left\{\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} > c\right\}. \quad \text{(A.8)}$$

Taking into account that $c > \Omega(f)$ and introducing the notation $\mu = c - \Omega(f)$, we obtain from (A.8) the assertion of the theorem:

$$P\{\rho_1(f_l^{\gamma_l}, f) > v\} \leq P\{\rho_2^2(F_l, F) > \mu\gamma_l\}. \quad \text{(A.9)}$$

□

PROOF OF THEOREM A.2

(1) An arbitrary closed sphere in a Hilbert space (i.e., a set of vectors of the form $\{f: \|f - f_0\| \leq d\}$) is weakly compact. Therefore, as far as the weak compactness in the space E_1 is concerned, we are under the conditions of

Theorem A.1. Consequently, for any positive v and μ there exists a number $n = n(\mu, v)$ such that for $l > n(\mu, v)$

$$P\{|\varphi(f_l^{\gamma_l}) - \varphi(f)| > v\} \leq P\{\rho_2^2(F_l, F) > \gamma_l \mu\},$$

where $\varphi(\cdot)$ is an arbitrary continuous linear functional, for example the projection of f on the element q:

$$\varphi(f) = \int q(t) f(t)\, dt = (q \cdot f).$$

(2) According to the definition of a norm in a Hilbert space we have

$$\begin{aligned} \|f_l^{\gamma_l} - f\|^2 &= (f_l^{\gamma_l} - f, f_l^{\gamma_l} - f) \\ &= \|f_l^{\gamma_l}\|^2 - \|f\|^2 + 2(f, f - f_l^{\gamma_l}). \end{aligned} \tag{A.10}$$

Utilizing the inequality

$$P\{a + b > \varepsilon\} \leq P\left\{a > \frac{\varepsilon}{2}\right\} + P\left\{b > \frac{\varepsilon}{2}\right\},$$

we obtain from (A.10) that

$$P\{\|f_l^{\gamma_l} - f\|^2 > \varepsilon\} \leq P\left\{\|f_l^{\gamma_l}\|^2 - \|f\|^2 > \frac{\varepsilon}{2}\right\} + P\left\{2(f, f - f_l^{\gamma_l}) > \frac{\varepsilon}{2}\right\}.$$

In order to bound the first summand on the right-hand side we shall utilize the inequality (A.3a), taking into account that $\Omega(f) = \|f\|^2$. We thus obtain

$$\|f_l^{\gamma_l}\|^2 \leq \|f\|^2 + \frac{\rho_2^2(F_l, F)}{\gamma_l}.$$

Therefore

$$P\left\{\|f_l^{\gamma_l}\|^2 - \|f\|^2 > \frac{\varepsilon}{2}\right\} \leq P\left\{\frac{\rho_2^2(F_l, F)}{\gamma_l} > \frac{\varepsilon}{2}\right\}. \tag{A.11}$$

We bound the second summand by means of (A.9), setting $\mu = \varepsilon/2$

$$P\left\{(f, f - f_l^{\gamma_l}) > \frac{\varepsilon}{4}\right\} \leq P\left\{\rho^2(F_l, F) > \frac{\varepsilon}{2}\gamma_l\right\}. \tag{A.12}$$

Combining the bounds (A.11) and (A.12), we arrive at the assertion of the theorem:

$$P\{\|f_l^{\gamma_l} - f\|^2 > \varepsilon\} \leq 2P\left\{\rho_2^2(F_l, F) > \frac{\varepsilon}{2}\gamma_l\right\}. \qquad \square$$

Chapter 10

Estimation of Functional Values at Given Points

§1 The Scheme of Minimizing the Overall Risk

In the case of small samples

$$x_1, y_1; \ldots; x_l, y_l \tag{10.1}$$

it is desirable to distinguish between two estimation problems:

(1) estimation in the class $F(x, \alpha)$ of the functional dependence $y = f(x)$,
(2) estimation in $F(x, \alpha)$ of values of a function $y = f(x)$ at the given points

$$x_{l+1}, \ldots, x_{l+k}. \tag{10.2}$$

It may seem that the problem of estimating the values of a function $y = f(x)$ at given points (10.2) is not a very profound one. There exists a "natural" way to solve it: based on the available empirical data (10.1), estimate the functional dependence $y = F(x, \alpha^*)$, and using this estimate determine the values of the function at the points (10.2):

$$y_i = F(x_i, \alpha^*) \qquad (i = l + 1, \ldots, l + k),$$

i.e., one can obtain a solution of the second problem by using a solution of the first one. However such a route for estimating values of a function is often not the best, since here a solution of a relatively simple problem—estimating k numbers (the values of the function) becomes dependent on the solution of a substantially more complex problem—estimating a function (which is a continuum containing these k numbers).

The problem is actually how to utilize the information under the conditions of incompleteness of information—for solving the required problem rather than a more general one. It may turn out that the amount of information available will be sufficient to estimate the k numbers satisfactorily,

but not be sufficient to estimate the function in the whole domain of its definition.

It should be noted that in practice usually the need arises to determine the values of the functions at given points rather than to determine the functional dependence itself. As a rule (which is always valid for the problem of pattern recognition), the functional dependence is utilized only to determine the values of a function at certain desired points.

Thus we distinguish between two kinds of estimation problems: *estimation of a function* and *estimation of values of a function at given points.*

In Chapter 1 we formalized the statement of the problem of estimation of functional dependence by means of a scheme of minimizing the expected risk. In this section we shall formalize the statement of the problem of estimating functional values at given points using a scheme which will be called the scheme of *minimizing the overall risk.*

It is assumed that a set

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k}, \tag{10.3}$$

consisting of $l + k$ vectors (a *complete sample of vectors*) is given. There exists a function $y = f(x)$ which assigns a number y to each vector x in the set (10.3). Thus for $l + k$ vectors (10.3), $l + k$ values

$$y_1, \ldots, y_l, y_{l+1}, \ldots, y_{l+k} \tag{10.4}$$

are defined. From the set (10.3) l vectors x_i are randomly selected for which the corresponding realizations of y_i are indicated. The set of pairs

$$x_1, y_1; \ldots; x_l, y_l \tag{10.5}$$

thus formed will be called the *training sample.* The set of vectors

$$x_{l+1}, \ldots, x_{l+k} \tag{10.6}$$

is called the *working sample.*

Based on the elements of the training and working samples and on a given set of functions $F(x, \alpha)$ ($f(x)$ does not necessarily belong to this set), it is required to obtain a function $F(x, \alpha^*)$ that minimizes with a preassigned probability $1 - \eta$ the overall risk of forecasting the values of the function $y_i = f(x_i)$ on the elements of the working sample, i.e., which yields with probability $1 - \eta$ a value of the functional

$$I_\Sigma(\alpha) = \frac{1}{k} \sum_{i=l+1}^{l+k} (y_i - F(x_i, \alpha))^2 \tag{10.7}$$

close to the minimal one.

We shall call this formulation of the problem of estimating the values of a function at given points *formulation I*, and we shall consider another formulation of this problem, to be referred to as *formulation II.*

Let a probability distribution $P(x, y)$ be given on the set of pairs (X, Y). We select from this set, randomly and independently, l pairs

$$x_1, y_1; \ldots; x_l, y_l,$$

which form the training sequence. Next, in the same manner, k additional pairs

$$x_{l+1}, y_{l+1}; \ldots; x_{l+k}, y_{l+k}$$

are chosen.

It is required to obtain an algorithm A which, based on the training sequence $x_1, y_1; \ldots; x_l, y_l$ and the working sample $x_{l+1}, \ldots, x_{l+k}$, will choose in $F(x, \alpha)$ a function

$$F(x, \alpha_A(x_1, y_1; \ldots; x_l, y_l; x_{l+1}, \ldots, x_{l+k}))$$

which yields a value of the functional

$$I_r(A) = \int \frac{1}{k} \sum_{i=l+1}^{l+k} (y_i - F(x_i, \alpha_A(x_1, y_1; \ldots; x_l, y_l; x_{l+1}, \ldots, x_{l+k})))^2 \times P(x_1, y_1) \cdots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \cdots dx_{l+k}\, dy_{l+k}$$

close to the minimal one.

The following theorem which connects these two formulations is valid.

Theorem 10.1. *If for some algorithm A it is proved that for formulation I with probability $1 - \eta$ the deviation between the risk on the training and working samples does not depend on the composition of the complete sample and does not exceed $\varkappa$, then with the same probability for formulation II the deviation between the analogous values of the risks does not exceed $\varkappa$.*

PROOF. Denote

$$C_A(x_1, y_1; \ldots; x_{l+k}, y_{l+k}) = \left| \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha_A))^2 - \frac{1}{k} \sum_{i=l+1}^{l+k} (y_i - F(x_i, \alpha_A))^2 \right|.$$

Consider the second formulation of the problem, and compute the probability of deviation from zero by an amount greater than $\varkappa$ of the quantity $C_A(x_1, y_1, \ldots; x_{l+k}, y_{l+k})$:

$$P = \int_{XY} \theta[C_A(x_1, y_1; \ldots; x_{l+k}, y_{l+k}) - \varkappa] \times P(x_1, y_1) \cdots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \cdots dx_{l+k}\, dy_{l+k},$$

where

$$\theta(z) = \begin{cases} 1 & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

Let T_p $(p = 1, 2, \ldots (l + k)!)$ be the permutation operator for the sample $x_1, y_1, \ldots, x_{l+k}, y_{l+k}$. Then the equality

$$\begin{aligned} P &= \int_{XY} \theta[C_A(x_1, y_1; \ldots; x_{l+k}, y_{l+k}) - \varkappa] \\ &\quad \times P(x_1, y_1) \cdots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \cdots dx_{l+k}\, dy_{l+k} \\ &= \int_{XY} \left\{ \frac{1}{(l+k)!} \sum_{p=1}^{(l+k)!} \theta[C_A(T_p(x_1, y_1; \ldots; x_{l+k}, y_{l+k})) - \varkappa] \right\} \\ &\quad \times P(x_1, y_1) \cdots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \cdots dx_{l+k}\, dy_{l+k} \end{aligned}$$

is valid. The expression in braces is the quantity estimated in formulation I. It does not exceed $1 - \eta$. We thus obtain

$$P \leq \int_{XY} (1 - \eta) P(x_1, y_1) \cdots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \cdots dx_{l+k}\, dy_{l+k} = 1 - \eta.$$

The theorem is proved. □

Below we shall consider the problem of estimating values of a function at given points in formulation I. However, by means of Theorem 10.1 all the results obtained can be shown to be valid for the case of formulation II as well.

In this chapter the terminology used pertains to estimating values of a function. However, all the results obtained were valid in the more general case when a realization of the sample (10.4) is determined by the conditional probability $P(y|x)$ (rather than by the function $y = f(x)$) and it is required on the basis of random realizations at points (10.5) to forecast, by means of a function belonging to $F(x, \alpha)$, realizations at some other points (10.6).

§2 The Method of Structural Minimization of the Overall Risk

We solve the problem of estimating values of a function at given points using the method of structural risk minimization. In the following two sections we obtain bounds on the rate of uniform relative deviation of the mean values in two subsamples. Using these bounds, we construct bounds on the overall risk, uniform over the class $F(x, \alpha)$, based on the values of the empirical risks. These bounds are analogous to those which were utilized in the preceding chapters when constructing a structural minimization of the expected risk.

We shall demonstrate that for a set of indicator functions of capacity h (in the problem of pattern recognition) the bound

$$\nu_\Sigma(\alpha) < \nu(\alpha) + \Omega_1^*(l, k, h, -\ln \eta) \tag{10.8}$$

is valid with probability $1 - \eta$ (for this problem the notation $I_\Sigma(\alpha) = \nu_\Sigma(\alpha)$, $I_{\text{emp}}(\alpha) = \nu(\alpha)$ is used), while for the set of arbitrary functions of capacity h, with probability $1 - \eta$ the bound of the form

$$I_\Sigma(\alpha) < I_{\text{emp}}(\alpha)\Omega_2^*(l, k, h, -\ln \eta) \tag{10.9}$$

is valid.

Now if one defines the structure

$$S_1 \subset \cdots \subset S_q$$

on the class of functions $F(x, \alpha)$, then it is possible by minimizing the right-hand side of the inequality (10.8) (or (10.9)) to find an element S_* and a

function $F(x, \alpha^*_{\text{emp}})$ for which the guaranteed minimum for the bound of the overall risk is attained. Using the functions $F(x, \alpha^*_{\text{emp}})$, the values $y_i = F(x_i, \alpha^*_{\text{emp}})$ are computed at the points of the working sample. Outwardly this scheme does not differ at all from the one considered in Chapter 8.

However, in the scheme of structural minimization of the overall risk there is a special feature which determines the difference between solutions of problems of estimating functions and those of estimating values of a function at given points. This has to do with the need to order the functions in the class $F(x, \alpha)$ *a priori*. This requirement has different meanings in the cases of estimating functions and of estimating values of functions. For the problem of estimating functions it means that, knowing the class of functions $F(x, \alpha)$ and the domain of definition of a function, it is necessary to define a structure on $F(x, \alpha)$. For the problem of estimating functional values it amounts to determining a structure on $F(x, \alpha)$, knowing the class of functions $F(x, \alpha)$ and the complete sample

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k}. \tag{10.10}$$

The difference is that for a complete sample (10.10) the set of functions $F(x, \alpha)$ is decomposed into sets of *equivalence classes*. This set can be investigated, and the structure on $F(x, \alpha)$ can be defined on equivalence classes, producing a more meaningful ordering principle than the one in the case of estimating functions.

For example, the set of indicator functions on the complete sample (10.10) is decomposed into a finite number of equivalence classes. Two indicator functions are equivalent on a complete sample if they subdivide this sample into subsamples in the same manner (i.e., take the same values on (10.10)). In this case it makes sense to define a structure on a finite number of equivalence classes rather than on the initial set of functions.

Below, when discussing estimation of values of functions at given points, we shall consider three different conceptions of defining and ordering equivalence classes, and each one of them will be implemented to estimate values of indicator functions as well as to estimate values of arbitrary functions. First, however, we shall obtain bounds which serve as the basis for the method of structural minimization of the overall risk.

§3 Bounds on the Uniform Relative Deviation of Frequencies in Two Subsamples

In this section we shall prove a theorem on the uniform relative deviation of frequencies in two subsamples. For the problem of minimizing the overall risk in the class of indicator functions this theorem plays the same role as the theorem on uniform relative deviations of frequencies from their proba-

bilities played in the problem of minimizing the expected risk. To state the theorem we shall introduce the function $\Gamma_{l,k}(\varkappa)$.

Let the set

$$x_1, \ldots, x_{l+k}$$

be given, consisting of elements of two types: m elements of type a and $l + k - m$ elements of type b. We select randomly l elements from this set. The probability that among the selected elements there will be r elements of type a equals

$$P(r, l + k, l, m) = \frac{C_m^r C_{l+k-m}^{l-r}}{C_{l+k}^l}. \tag{10.11}$$

Thus with probability (10.11) the frequency of elements of type a in the selected group is r/l, and hence the corresponding frequency in the remaining group will be $(m - r)/k$.

The probability that the frequency of elements a in the first group will deviate from the frequency of elements a in the second by the amount exceeding $\varkappa$ is equal to

$$P\left\{\left|\frac{r}{l} - \frac{m-r}{k}\right| > \varkappa\right\} = \sum_r \frac{C_m^r C_{l+k-m}^{l-r}}{C_{l+k}^l} = \Gamma_{l,k}(\varkappa, m),$$

where the summation is taken over the values of r such that

$$\left|\frac{r}{l} - \frac{m-r}{k}\right| > \varkappa, \qquad \max(0, m - k) \leq r \leq \min(m, l).$$

We define the function

$$\Gamma_{l,k}(\varkappa) = \max_m \Gamma_{l,k}\left(\sqrt{\frac{m}{l+k}}\,\varkappa, m\right).$$

This function can easily be tabulated with a computer.

Denote now by $\nu_0(\alpha)$ the frequency of errors incurred in the classification of the set $x_1, \ldots, x_{l+k}$ when using the decision rule $F(x, \alpha)$. Clearly

$$\nu_0(\alpha) = \frac{k}{l+k}\,\nu_\Sigma(\alpha) + \frac{l}{l+k}\,\nu(\alpha). \tag{10.12}$$

The following theorem on uniform relative deviation of frequencies in the two subsamples is valid.

Theorem 10.2. *Let the class of decision rules $F(x, \alpha)$ possess the capacity $h < l + k$. Then the probability that the relative size of the deviation for at least one rule in $F(x, \alpha)$ exceeds $\varkappa$ is bounded by*

$$P\left\{\sup_\alpha \frac{|\nu(\alpha) - \nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}} > \varkappa\right\} < 1.5\,\frac{(l+k)^h}{h!}\,\Gamma_{l,k}(\varkappa). \tag{10.13}$$

Here we use the convention $|\nu(\alpha) - \nu_\Sigma(\alpha)|/\sqrt{\nu_0(\alpha)} = 0$ for $\nu(\alpha) = \nu_\Sigma(\alpha) = \nu_0(\alpha) = 0$.

PROOF. Observe that the number of equivalence classes on the complete samples does not exceed $N = m^S(l + k)$. Therefore the inequality

$$P\left\{\sup_{\alpha} \frac{|\nu(\alpha) - \nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}} > \varkappa\right\} < N \sup_{\alpha} P\left\{\frac{|\nu(\alpha) - \nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}} > \varkappa\right\}$$

is valid. For $h < l + k$ the first term on the right-hand side is bounded by $1.5(l + k)^h/h!$, while the second term is bounded by the function $\Gamma_{l,k}(\varkappa)$. Indeed,

$$\begin{aligned} P\left\{\frac{|\nu(\alpha) - \nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}} > \varkappa\right\} &= P\{|\nu(\alpha) - \nu_\Sigma(\alpha)| > \varkappa\sqrt{\nu_0(\alpha)}\} \\ &= P\left\{|\nu(\alpha) - \nu_\Sigma(\alpha)| > \varkappa\sqrt{\frac{m.}{l + k}}\right\} \\ &= \Gamma_{l,k}\left(\varkappa\sqrt{\frac{m}{l + k}}, m\right), \end{aligned}$$

and by definition

$$\Gamma_{l,k}\left(\sqrt{\frac{m}{l + k}}\varkappa, m\right) \leq \Gamma_{l,k}(\varkappa).$$

The theorem is proved. □

Below, a bound uniform in $F(x, \alpha)$ on the frequency of errors in the working sample will be required. We shall derive it using Theorem 10.2. We bound the right-hand side of (10.13) by the quantity η. We thus arrive at the inequality

$$h\left(\ln\frac{l + k}{1} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leq \ln\frac{\eta}{1.5},$$

the smallest solution of which will be denoted by $\varkappa_*$.

Taking (10.12) into account, we obtain from (10.13) that with probability $1 - \eta$ the inequality

$$\nu_\Sigma(\alpha) < \nu(\alpha) + \frac{k\varkappa_*^2}{2(l + k)} + \varkappa_*\sqrt{\nu(\alpha) + \left(\frac{k\varkappa_*}{2(l + k)}\right)^2} \tag{10.14}$$

is valid for all α. We shall utilize this inequality when constructing algorithms for structural minimization of the risk in the class of indicator functions.

§4 A Bound on the Uniform Relative Deviation of Means in Two Subsamples

When deriving a bound on the uniform relative deviation of the means in two subsamples we shall assume that on the complete sample

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k} \tag{10.15}$$

the condition

$$\sup_{\alpha} \frac{\sqrt[p]{\frac{1}{l+k}\sum_{i=1}^{l+k}(y_i - F(x_i, \alpha))^{2p}}}{\frac{1}{l+k}\sum_{i=1}^{l+k}(y_i - F(x_i, \alpha))^2} \leq \tau \qquad (p > 2) \tag{10.16}$$

is fulfilled for a set of arbitrary functions $F(x, \alpha)$, where y_i is a value of the realization of (10.4).

The condition (10.16) conveys some prior information concerning possible large deviations on the complete sample (10.15). This condition is analogous to the condition considered in Section 6 of Chapter 7.

In the same manner as in Chapter 7, we introduce the function

$$R_1(\alpha) = \int \sqrt{v((y - F(x, \alpha))^2 > t)}\, dt,$$

where $v\{(y - F(x, \alpha))^2 > t\}$ is the ratio of the number of points in the complete sample (10.15) for which the condition $(y - F(x, \alpha))^2 > t$ is fulfilled on realizations of (10.4) to the total number of points $l + k$. For the function $R_1(\alpha)$, as for the function $R(\alpha)$ (cf. Chapter 7), the relation

$$R_1(\alpha) < a(p) \sqrt[p]{\frac{1}{l+k}\sum_{i=1}^{l+k}(y_i - F(x_i, \alpha))^{2p}} \tag{10.17}$$

is fulfilled, where

$$a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}.$$

Denote

$$I(\alpha) = \frac{1}{l+k}\sum_{i=1}^{l+k}(y_i - F(x_i, \alpha))^2$$

$$= \frac{1}{l+k} I_{\text{emp}}(\alpha) + \frac{k}{l+k} I_{\Sigma}(\alpha). \tag{10.18}$$

The following theorem is valid.

Theorem 10.3. *Let the condition* (10.16) *be fulfilled and the class of functions possess capacity $h < l + k$. Then the bound*

$$P\left\{\sup_{\alpha} \frac{|I_{\Sigma}(\alpha) - I_{\text{emp}}(\alpha)|}{I(\alpha)} > \tau a(p) \varkappa\right\} < 1.5 \frac{(l+k)^h}{h!} \Gamma_{l,k}(\varkappa) \tag{10.19}$$

is valid.

PROOF. To prove the theorem we shall utilize the assertion of Theorem 10.2 according to which the bound

$$P\left\{\sup_{\alpha}\frac{|\nu(A_{\alpha,\beta})-\nu_{\Sigma}(A_{\alpha,\beta})|}{\sqrt{\nu_0(A_{\alpha,\beta})}}>\varkappa\right\}<1.5\frac{(l+k)^h}{h!}\Gamma_{l,k}(\varkappa) \tag{10.20}$$

is valid. (Here $\nu(A_{\alpha,\beta})$ is the frequency of the event $\{(y-F(x,\alpha))^2>\beta\}$ computed for the training sequence, $\nu_{\Sigma}(A_{\alpha,\beta})$ is the frequency of the event $\{(y-F(x,\alpha))^2>\beta\}$ computed for the working sample via the realization of (10.4), and $\nu_0(A_{\alpha,\beta})$ is the frequency of $\{(y-F(x,\alpha))^2>\beta\}$ computed for the complete sample (10.15) via the realization of (10.4).)

We shall show that the validity of (10.20) implies the validity of the inequality

$$P\left\{\sup_{\alpha}\frac{|I_{\Sigma}(\alpha)-I_{\text{emp}}(\alpha)|}{R_1(\alpha)}>\varkappa\right\}<1.5\frac{(l+k)^h}{h!}\Gamma_{l,k}(\varkappa). \tag{10.21}$$

For this purpose we write the expression

$$\sup_{\alpha}\frac{|I_{\Sigma}(\alpha)-I_{\text{emp}}(\alpha)|}{R_1(\alpha)}\le\sup_{\alpha}\lim_{n\to\infty}R$$

in the form of a Lebesgue integral, where

$$R=\frac{\sum_{i=1}^{\infty}\frac{1}{n}\left|\nu_{\Sigma}\left\{(y-F(x,\alpha))^2>\frac{i}{n}\right\}-\nu_{\text{emp}}\left\{(y-F(x,\alpha))^2>\frac{i}{n}\right\}\right|}{R_1(\alpha)}.$$

Now let the inequality

$$\frac{\left|\nu_{\Sigma}\left\{(y-F(x,\alpha))^2>\frac{i}{n}\right\}-\nu_{\text{emp}}\left\{(y-F(x,\alpha))^2>\frac{i}{n}\right\}\right|}{\sqrt{\nu\left\{(y-F(x,\alpha))^2>\frac{i}{n}\right\}}}\le\varkappa$$

be valid. In that case

$$\frac{|I_{\Sigma}(\alpha)-I_{\text{emp}}(\alpha)|}{R_1(\alpha)}\le\lim_{n\to\infty}\frac{\varkappa\sum_{i=1}^{\infty}\frac{1}{n}\sqrt{\nu\left\{(y-F(x,\alpha))^2>\frac{i}{n}\right\}}}{R_1(\alpha)}=\varkappa.$$

The the validity of (10.20) implies that (10.21) holds.

To complete the proof it remains only to utilize the inequalities (10.17) and (10.21). Indeed

$$P\left\{\sup_{\alpha}\frac{|I_{\Sigma}(\alpha)-I_{\text{emp}}(\alpha)|}{I(\alpha)}>\tau a(p)\varkappa\right\}\le P\left\{\sup_{\alpha}\frac{|I_{\Sigma}(\alpha)-I_{\text{emp}}(\alpha)|}{R_1(\alpha)}>\varkappa\right\}$$
$$\le 1.5\frac{(l+k)^h}{h!}\Gamma_{l,k}(\varkappa).$$

The theorem is proved. □

We shall now obtain a uniform bound for the risk on the working sample. For this purpose we bound the right-hand side of (10.19) by a quantity η. We thus arrive at the inequality

$$h\left(\ln\frac{l+k}{h}+1\right)+\ln\Gamma_{l,k}(\varkappa)\leq\ln\frac{\eta}{1.5}. \tag{10.22}$$

Denote by $\varkappa_*$ the smallest solution for this inequality.

Taking the representation (10.18) into account, we obtain from (10.19) that the inequality

$$I_\Sigma(\alpha)<\left[\frac{1+\tau a(p)\dfrac{l}{l+k}\varkappa_*}{1-\tau a(p)\dfrac{k}{l+k}\varkappa_*}\right]_\infty I_{\text{emp}}(\alpha), \tag{10.23}$$

where

$$[z]_\infty=\begin{cases}z & \text{for } z\geq 0,\\ \infty & \text{for } z<0\end{cases}$$

is valid with probability $1-\eta$.

This inequality will be utilized in the course of constructing algorithms for a structural minimization of the overall risk. Below we shall confine ourselves to a class of functions linear in parameters,

$$F(x,\alpha)=\sum_{i=1}^{n-1}\alpha_i\varphi_i(x)+\alpha_0.$$

The capacity of this class of functions equals n.

§5 Estimation of Values of an Indicator Function in a Class of Linear Decision Rules

Let a complete sample

$$x_1,\ldots,x_i,x_{l+1},\ldots,x_{l+k} \tag{10.24}$$

be given. On this sample the set of decision rules is decomposed into a finite number N of equivalence classes $F_1,\ldots,F_N$. Two decision rules $F(x,\hat{\alpha})$ and $F(x,\hat{\hat{\alpha}})$ fall into the same equivalence class if they subdivide the sample (10.24) into two subsamples in the same manner. Altogether $\Delta^S(x_1,\ldots,x_{l+k})$ subdivisions of the sample (10.24) into two classes by means of the rules $F(x,\alpha)$ are possible, and thus there exist $\Delta^S(x_1,\ldots,x_{l+k})$ equivalence classes.

In view of the definition (cf. Chapter 6, Section 7),

$$\Delta^S(x_1,\ldots,x_l)\leq m^S(l).$$

For linear decision rules in a space of dimension n the following bound is valid (cf. Chapter 6, Section 8):

$$m^S(l + k) < 1.5 \frac{(l + k)^n}{n!}.$$

Thus on the complete sample (10.24) the set of linear decision rules $F(x, \alpha)$ is decomposed into $N \leq 1.5(l + k)^n/n!$ equivalence classes $F_1, \ldots, F_N$.

Observe that the equivalence classes are not of equal size. Some of them contain more decision rules than others. We assign to each equivalent class a quantity which will characterize the fraction of linear decision rules belonging to this class relative to all linear decision rules. Such a quantity can be constructed. Indeed, assign to each function

$$F(x, \alpha) = \theta\left(\sum_{i=1}^{n} \alpha_i \varphi_i(x)\right)$$

a directional vector (Figure 19)

$$\alpha = (\alpha_1, \ldots, \alpha_n)^T; \qquad \|\alpha\| = 1.$$

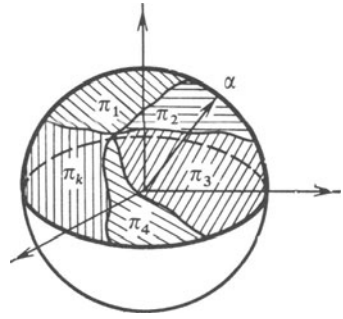

Figure 19

Then in the space of parameters α a unit sphere corresponds to the set of all hyperplanes; and to each equivalence class F_i there corresponds a distinct region on the surface of the sphere. (The set of N equivalence classes subdivides the sphere into N regions.) The ratio of the area corresponding to the region $\mathscr{L}_i$ to the area of the sphere $\mathscr{L}$ characterizes the fraction of functions belonging to an equivalence class relative to all possible linear decision rules.

We now order the equivalence classes in decreasing order of $\pi_i = \mathscr{L}_i/\mathscr{L}$ and introduce the following structure:

$$S_1 \subset S_2 \subset \cdots \subset S_q, \tag{10.25}$$

where the element S_p contains only those equivalence classes which satisfy

$$\frac{\mathscr{L}_i}{\mathscr{L}} > c_p \qquad (c_1 > c_2 > \cdots > c_q = 0).$$

We have thus constructed a structure in which each element S_p possesses an extremal property: for a given number of equivalence classes it contains the maximal share of all decision rules. However, in practice it is troublesome to compute the values $\mathscr{L}_i/\mathscr{L}$ and thus to form the structure (10.25). We shall therefore consider another characteristic of the size of an equivalence class which is similar to $\mathscr{L}_i/\mathscr{L}$ in its meaning and can be obtained in practice.

Denote by ρ_p the value of the distance between the convex hulls of vectors of the complete sample allocated to different classes by the decision rules belonging to F_p, and assign to the equivalence class F_p the number

$$\pi(F_p) = \frac{\rho_p}{D}, \tag{10.26}$$

where $D/2$ is the radius of the minimal sphere containing the set (10.24), i.e.,

$$\frac{D}{2} = \min_{x^*} \max_{x_1, \ldots, x_{l+k}} \|x_i - x^*\|.$$

Now define a structure

$$S_1 \subset S_2 \subset \cdots \subset S_n \tag{10.27}$$

on the equivalence classes; here S_d contains only those equivalence classes F_i such that

$$\begin{aligned} &\pi^2(F_i) > \frac{1}{d-1} && \text{for } d < n, \\ &\pi^2(F_i) \geq 0 && \text{for } d = n,\ d \geq 2. \end{aligned} \tag{10.28}$$

The set S_1 in (10.27) is empty.

To construct a method of structural minimization for the overall risk on the structure (10.27) we shall bound the number N_d of equivalence classes belonging to the elements of the structure S_d.

The following lemma is valid.

Lemma. *The number N_d of equivalence classes in S_d is bounded by*

$$N_p < 1.5 \frac{(l+k)^d}{d!}, \tag{10.29}$$

where

$$d = \min\left(n, \left[\frac{D^2}{\rho^2}\right] + 1\right), \tag{10.30}$$

n is the dimensionality of the space, and $[a]$ is the integral part of number a.

Proof. Observe that the number N_d equals the maximal number of subdivisions of the sample

$$x_1, \ldots, x_{l+k}$$

into two classes such that the distance between their convex hulls exceeds $D/\sqrt{d-1}$, i.e.,

$$\rho > \frac{D}{\sqrt{d-1}} = \rho_d. \tag{10.31}$$

The number of such decision rules does not exceed

$$m^S(l+k) < 1.5\,\frac{(l+k)^r}{r!},$$

where r is the maximal number of points in the sample for which an arbitrary subdivision into two classes satisfies (10.31). Observe that if the condition (10.31) is fulfilled, then the subdivision is evidently carried out by means of a hyperplane; therefore obviously

$$r \le n,$$

where n is the dimension of the space.

Now consider r points

$$x_1, \ldots, x_r$$

and 2^r possible subdivisions of these points into two subsets

$$T_1, \ldots, T_{2^r}.$$

Denote by $\rho_p(T_i)$ the distance between the convex hulls of vectors belonging to distinct subsets under subdivision T_i.

The fact that (10.31) is fulfilled for any T_i can be written as

$$\min_i \rho(T_i) > \rho_d.$$

Then the number r does not exceed the maximal number of vectors such that the inequality

$$H(r) = \max_{x_1,\ldots,x_r} \min_i \rho(T_i) \ge \frac{D}{\sqrt{d-1}}\,\rho_d \tag{10.32}$$

is still fulfilled. It follows from symmetry considerations that the maximal r is attained where the vectors $x_1, \ldots, x_r$ are located at the vertices of a regular $(r-1)$-dimensional simplex inscribed in a sphere of radius $D/2$, and T_i is a subdivision into two subsimplices of dimension $(r/2)-1$ for even r, and two subsimplices of dimensions $(r-1)/2$ and $(r-3)/2$ for odd r. Therefore elementary calculations show that

$$H(r) = \begin{cases} \dfrac{D}{\sqrt{r-1}} & \text{for even } r, \\[2ex] \dfrac{D}{\sqrt{r-1}}\sqrt{\dfrac{r^2}{r^2-1}} & \text{for odd } r. \end{cases}$$

For $r \geq 10$ the quantities

$$\frac{1}{\sqrt{r-1}} \quad \text{and} \quad \sqrt{\frac{r^2}{r^2-1}} \cdot \frac{1}{\sqrt{r-1}}$$

are close to each other (they differ by an amount less than 0.01). Thus we take

$$H(r) = \frac{D}{\sqrt{r-1}}. \tag{10.33}$$

(A bound from the above on $H(r)$ would have been the expression

$$H(r) \leq \frac{D}{\sqrt{r-1.01}} \qquad (r > 10).\Big)$$

It follows from the inequalities (10.32) and (10.33) that for integer r

$$r < \left[\frac{D^2}{\rho^2}\right] + 1.$$

Finally, taking into account that the subdivision is done by means of a hyperplane, i.e., $r \leq n$, we obtain

$$d \leq \min\left(\left[\frac{D^2}{\rho^2}\right] + 1, n\right). \tag{10.34}$$

Consequently in view of Theorem 6.6 we have

$$N_d < 1.5\frac{(l+k)^d}{d!}.$$

The lemma is thus proved. □

It follows from Theorem 10.2 and the lemma that with probability $1 - \eta$ simultaneously for all decision rules in S_d the inequality

$$v_\Sigma(\alpha) < v(\alpha) + \frac{k\varkappa_*^2}{2(l+k)} + \varkappa_3\sqrt{v(\alpha) + \left(\frac{k\varkappa_*}{2(l+k)}\right)^2} = R(\alpha, d) \tag{10.35}$$

is fulfilled, where $\varkappa_*$ is the smallest solution of the inequality

$$d\left(\ln\frac{l+k}{d} + 1\right) + \ln\Gamma_{l,k}(\varkappa) \leq \ln\frac{\eta}{1.5}.$$

The method of structural minimization of the overall risk consists of indexing the working sample by means of the rule $F(x, \alpha^*_{\text{emp}})$ which minimizes the functional (10.35) with respect to d and α. Let the minimum be equal to $R(\alpha^*_{\text{emp}}, d_*)$. For such an indexing procedure the assertion

$$P\{v_\Sigma(\alpha^*_{\text{emp}}) < R(\alpha^*_{\text{emp}}, d_*)\} > 1 - n\eta$$

is valid.

In Addendum 1 we shall present a description of algorithms which minimize the overall risk in the class of linear decision rules. Here we shall consider an example which illustrates the difference between solving the problem of classifying vectors in a working sample using the method of minimizing the overall risk and using a decision rule which minimizes the empirical risk for a training sequence.

In Figure 20, vectors of the first class of the training sequence are denoted by crosses, and vectors of the second class by small circles. Dots represent vectors of the trial sample.

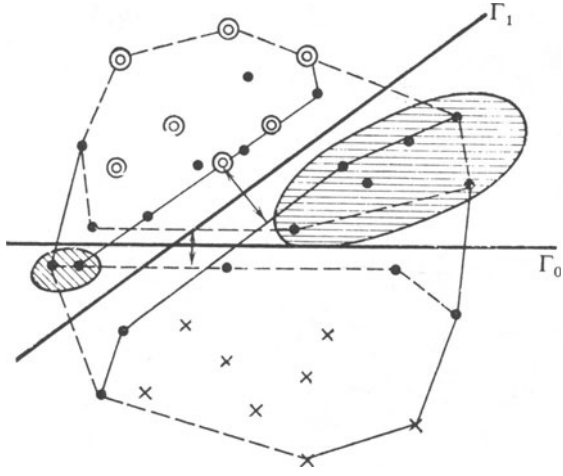

Figure 20

A solution of this problem within the framework of minimizing the expected risk consists in constructing a subdividing hyperplane which will assure the minimal probability of error. Let a solution be chosen among hyperplanes which errorlessly subdivides the vectors of the training sequence. In this case the minimal guaranteed probability of error is assured by the optimal subdividing hyperplane (the one that is the farthest from the elements of the training sequence). The vectors which are located on different sides of the hyperplane Γ_0 are assigned to different classes. This determines the solution of the problem using the method of minimizing the average risk. A solution of the problem using the method of minimizing the overall risk is determined by a hyperplane Γ_1 which maximizes the distance between the convex hulls of the subdivided sets. Vectors located on one side of the hyperplane belong to the first class, and those on the other side of the hyperplane belong to the second class.

Those points of the working sample which are classified by the hyperplanes Γ_0 and Γ_1 in a different manner are shaded in Figure 20.

§6 Selection of a Sample for Estimating Values of an Indicator Function

We have seen that the solution of the problem of estimating values of an indicator function at given points using the method of structural minimization of the overall risk leads to results which are different from those obtained from the classification of vectors of the working sample

$$x_{i+1}, \ldots, x_{l+k} \tag{10.36}$$

by means of a decision rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk on the elements of the training sequence

$$x_1, \omega_1; \ldots; x_l, \omega_l. \tag{10.37}$$

This result was obtained because the complete sample

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k} \tag{10.38}$$

consisted of a small number of elements whose location in the space could be studied; it is related to a specific method of ordering the class of decision rules $F(x, \alpha)$.

The method of ordering actually determined the difference in classification. Thus the geometry of vectors in the complete sample (10.38) predetermined the possibility of a more exact solution of the problem of estimating the values of a function at given points.

The method of ordering actually determined the difference in classification. Thus the geometry of vectors in the complete sample (10.38) predetermined the possibility of a more exact solution of the problem of estimating the values of a function at given points.

If this is indeed the case, then the question arises: is it possible, by excluding a few elements from the complete sample (10.38) (i.e., by changing the geometry of location of the vectors of the complete sample in space), to affect the definition of the structure on the class of decision rules in order to increase the guaranteed number of correct classifications of the elements in the working sample? It turns out that it is indeed possible.†

We shall now implement the idea of *selecting a complete sample.* Consider, along with the set X of vectors in the complete sample,

$$H_{l+k}^t = \sum_{p=0}^{t} C_{l+k}^p$$

distinct subsets $X_1, \ldots, X_{H_{l+k}^t}$ obtained from (10.38) by excluding at most t vectors. Now let a training sequence (10.37) and a working sample (10.36) be defined on the initial set of vectors (10.38). The training and working

† We note that in the case of estimating an indicator function the selection of the training sample does not lead to a decrease in the estimate of the minimal guaranteed risk.

samples induce on each one of the sets $X_1, \ldots, X_{H^t_{l+k}}$ its own training and working subsamples.

Consider H^t_{l+k} problems of estimating values of functions at given points. Each one of these problems is determined by a training sequence

$$x_1, \omega_1, \ldots, \widehat{x_i, \omega_i}, \ldots, \widehat{x_j, \omega_j}, \ldots, x_l, \omega_l$$

and a working sample

$$x_{l+1}, \ldots, \hat{x}_{l+\tau}, \ldots, x_{l+k}$$

($\hat{x}$ denotes that the element x is excluded from the sequence.)

For each problem, in accordance with its complete sample

$$x_1, \ldots, \hat{x}_i, \ldots, \hat{x}_j, \ldots, \hat{x}_{l+\tau}, \ldots, x_{l+k},$$

we shall determine equivalence classes of linear decision rules. We define a structure on the equivalence classes, utilizing the principle of ordering according to relative distances considered in the preceding section.

It follows from Theorem 10.2 and the lemma that with probability $1 - \eta$ in each problem (separately) the inequality

$$v_\Sigma(\alpha^d_{\text{emp}}) < v(\alpha^d_{\text{emp}}) + \frac{(k - k_{\text{ex}})\varkappa_*^2}{2(l + k - t_{\text{ex}})} + \varkappa_* \sqrt{v(a^d_{\text{emp}}) + \left[\frac{(k - k_{\text{ex}})\varkappa_*}{2(l + k - t_{\text{ex}}}\right]^2} \tag{10.39}$$

is valid for the rule $F(x, \alpha^d_{\text{emp}})$ minimizing the empirical risk in S_d, where $\varkappa_*$ is the smallest solution of the equation

$$d\left(\ln \frac{l + k - t_{\text{ex}}}{d} + 1\right) + \ln \Gamma_{l - l_{\text{ex}}, k - k_{\text{ex}}}(\varkappa) \leq \ln \frac{\eta}{1.5}. \tag{10.40}$$

In (10.39) and (10.40) the following notation is used: l_{ex} is the number of elements excluded from the training sequence, k_{ex} is the number of excluded elements from the working sample and $l_{\text{ex}} + k_{\text{ex}} = t_{\text{ex}}$.

Simultaneously for the dth elements of structures of all H^t_{l+k} problems the inequality

$$v^{(i)}_\Sigma(\alpha^d_{\text{emp}}) < v^{(i)}(\alpha^d_{\text{emp}}) + \frac{(k - k^{(i)}_{\text{ex}})}{2(l + k - t_i)}(\varkappa^{(i)}_*)^2 + \varkappa^{(i)}_* \sqrt{v^{(i)}(\alpha^d_{\text{emp}}) + \left[\frac{(k - k^{(i)}_{\text{ex}})\varkappa^{(i)}_*}{2(l + k - t_i)}\right]^2} \tag{10.41}$$

is fulfilled with probability $1 - \eta$, where $\varkappa^{(i)}_*$ are the smallest solutions of the inequalities

$$d\left(\ln \frac{l + k - t_i}{d} + 1\right) + \ln H^t_{l+k} + \ln \Gamma_{l - l^{(i)}_{\text{ex}}, k - k^{(i)}_{\text{ex}}}(\varkappa^{(i)}) \leq \ln \frac{\eta}{1.5}, \tag{10.42}$$

and i varies from 1 to H^t_{l+k}. In (10.41) and (10.42) the following notation is used: $l^{(i)}_{ex}$, $k^{(i)}_{ex}$ are the numbers of elements in the training and working samples omitted from (10.37) and (10.36) when forming the ith problem, $l^{(i)}_{ex} + k^{(i)}_{ex} = t_i$; $\nu^{(i)}_{\Sigma}(\alpha^d_{emp})$, $\nu^{(i)}(\alpha^d_{emp})$ are the frequencies of erroneous classification of the working and training samples in the ith problem.

Multiply each of the inequalities (10.41) by $k - k^{(i)}_{ex}$. This will yield for each problem a bound on the number of errors m_i in $k - k^{(i)}_{ex}$ elements of its working sample:

$$m_i < \left[\nu^{(i)}(\alpha^d_{emp}) + \frac{(k - k^{(i)}_{ex})}{2(l + k - t_i)}(\varkappa^{(i)}_*)^2 + \varkappa^{(i)}_* \sqrt{\nu^{(i)}(\alpha^d_{emp}) + \left[\frac{(k - k^{(i)}_{ex})}{2(l + k - t_i)}\varkappa^{(i)}_*\right]^2}\right](k - k^{(i)}_{ex}). \quad (10.43)$$

If the number of excluded vectors from the working sequences for all the problems were the same and equal to k_{ex}, then the best guaranteed solution of the problem of classifying $k - k_{ex}$ vectors in the working sample would be determined by the inequality (the problem) for which the value which bounds the number of errors in the $k - k_{ex}$ elements of the working sample is the smallest. However, the number of vectors excluded from the working sample is not the same for different problems. Therefore we shall consider as the best solution the one which maximizes the number of correct classifications of the elements of the working sample, i.e., is determined by the problem for which the minimum of the quantity†

$$R(d, i) = \left[\nu^{(i)}(\alpha^d_{emp}) + \frac{(k - k^{(i)}_{ex})}{2(l + k - t_i)}(\varkappa^{(i)}_*)^2 + \varkappa^{(i)}_* \sqrt{\nu^{(i)}(\alpha^d_{emp}) + \left(\frac{(k - k^{(i)}_{ex})\varkappa^{(i)}_*}{2(l + k - t_i)}\right)^2}\right](k - k^{(i)}_{ex}) + k^{(i)}_{ex} \quad (10.44)$$

(which determines the number of errors plus the number of excluded vectors from the working sample) is attained.

Now by enumeration over d and t we shall determine vectors which should be excluded in order to guarantee the largest number of correctly classified vectors in the working sample. The problem of minimizing the functional (10.44) with respect to d and t is quite difficult computationally. Its exact solution requires a large number of enumerations. Hosever, by using certain heuristic methods one can achieve a satisfactory solution in a reasonable amount of time. Details on algorithms for structural minimization of the overall risk are given in Addendum 1.

Observe that in the course of selection of a complete sample the elements of both the training sample and those of the working sample are picked.

† Here we can introduce different utilities (costs) for errors and refusal to classify elements in the working sample.

A selection of elements of the working sample allows us to increase the total number of correctly classified vectors at the expense of declining to classify certain elements.

Up until now we have assumed that the space on which the structure is constructed is fixed. However, the procedure of ordering with respect to relative distances may be carried out in any subspace E_m of the initial space E_n. Moreover, the minimal value of the corresponding bound need not be obtained on the initial space E_n. This fact yields the possibility of achieving a more refined minimum for the bound on the risk by means of additional minimization over subspaces.

§7 Estimation of Values of an Arbitrary Function in the Class of Functions Linear in Their Parameters

We shall now extend the methods of estimating values of indicator functions considered in the preceding sections to the estimation of values of an arbitrary function in a class of functions linear in their parameters. For this purpose we shall determine equivalence classes of linear (in parameters) functions on a complete sample, define a structure on these classes, and implement the method of structural risk minimization.

Let a complete sample

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k} \tag{10.45}$$

and a set of linear (in parameters) functions $F(x, \alpha)$ be given. We shall assign to each function $F(x, \alpha^*)$ in this set a one-parameter family (parameter β) of decision rules

$$F_{\alpha^*}(\beta) = \theta(F(x, \alpha^*) + \beta). \tag{10.46}$$

As the parameter β varies from $-\infty$ to ∞, the family (10.46) forms a sequence of dichotomies (subdivisions into two classes) of the set of vectors (10.45), starting with the dichotomy for which the first class is empty and the second class consists of the whole set of vectors (10.45),

$$[\varnothing; \{x_1, \ldots, x_{l+k}\}]$$

(for $\beta = -\infty$), and concluding with the dichotomy for which the first class contains the whole set (10.45) and the second class is empty,

$$[\{x_1, \ldots, x_{l+k}\}; \varnothing]$$

(for $\beta = +\infty$).

Thus for each function $F(x, \alpha)$ a sequence of dichotomies

$$[\varnothing, \{x_1, \ldots, x_{l+k}\}]; [\{x_{i_1}, \ldots, x_{i_l}\}; \{x_{j_1}, \ldots, x_{j_k}\}]; \ldots; [\{x_1, \ldots, x_{l+k}\}, \varnothing] \tag{10.47}$$

can be constructed. In accordance with this sequence of dichotomies we shall subdivide the set of functions $F(x, \alpha)$ into a finite number of equivalence classes. Two functions $F(x, \hat{\alpha})$ and $F(x, \hat{\hat{\alpha}})$ fall into the same equivalence class F_i if they form the same sequence of dichotomies (10.47).

Now assign to each equivalence class a number $\pi(F_i)$ which is equal to the fraction of all functions belonging to it, and then arrange the equivalence classes in the order of decreasing values of $\pi(F_i)$:

$$F_1, F_2, \ldots, F_N, \qquad \pi(F_1) \geq \pi(F_2) \geq \cdots \geq \pi(F_N). \tag{10.48}$$

Utilizing this ordering (10.48) one can construct a structure on the equivalence classes

$$S_1 \subset S_2 \subset \cdots \subset S_n.$$

The element S_r contains those equivalence classes for which $\pi(F_r) > c_r$.

One can define the fraction of functions belonging to an equivalence class in the case of sets of linear functions in the same manner as the fraction of linear decision rules was determined. Now assign to each linear function a vector of direction cosines. Then the surface of the unit sphere in the space of dimension n will correspond to the set of all functions, and a particular region on this sphere will correspond to each equivalence class (cf. Figure 19). The ratio of the area of a singled-out region to the area of the surface of the sphere will determine the fraction of functions belonging to an equivalence class among the whole set of functions.

In practice, however, it is difficult to compute the characteristic $\hat{\pi}(F_i)$ directly. Therefore, in the same manner as in Section 5, we shall consider another characteristic of the size of an equivalence class. For each function $F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x)$ we define a directional vector $\hat{\alpha} = \alpha/\|\alpha\|$. Each equivalence class F_m is characterized by the number

$$\rho_m = \min_{x_i, x_j} \sup_{\alpha} \left| (x_i - x_j)^{\mathrm{T}} \frac{\alpha}{\|\alpha\|} \right| \qquad (i \neq j),$$

where the minimum is taken over all the vectors of the complete sample and the supremum over all directional vectors of a given equivalence class.

We now form the following structure:

$$S_1 \subset \cdots \subset S_n.$$

The functions for which

$$\hat{\pi}^2(F) = \left[\frac{\rho}{D}\right]^2 > \frac{1}{d-1} \quad \text{for } d < n,$$

$$\hat{\pi}^2(F) = \left[\frac{\rho}{D}\right]^2 \geq 0 \qquad \text{for } d \geq n$$

—where D is the minimal diameter of the sphere containing the set $(x_1, \ldots, x_{l+k})$—are assigned to the dth element of the structure S_d. Utilizing

the lemma, one can show, as in Section 5, that the capacity of functions belonging to the S_dth element of the structure equals d, where

$$d = \min\left(\left[\frac{D^2}{\rho^2}\right] + 1, n\right).$$

The method of structural minimization for this structure involves finding an element S_* and a function $F(x, \alpha^*_{\text{emp}})$ in it such that the minimum on the right-hand side of the inequality

$$I_\Sigma(\alpha) < \left[\frac{1 + \tau a(p)\dfrac{l}{l+k}\varkappa_*}{1 - \tau a(p)\dfrac{k}{l+k}\varkappa_*}\right]_\infty I_{\text{emp}}(\alpha) \tag{10.49}$$

is obtained. Here $\varkappa_*$ is the smallest solution of the inequality

$$d\left(\ln\frac{l+k}{d} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leq \ln\frac{\eta}{1.5}.$$

The first factor on the right-hand side of (10.49) depends only on the order in which the vectors of the complete sample are projected on the vector of directions of the selected linear function, while the second factor depends on the value of the empirical risk.

Let the minimum of the right-hand side of (10.49) equal $R(\alpha^*_{\text{emp}}, d^*)$. Then the assertion

$$P\{I_\Sigma(\alpha^*_{\text{emp}}) < R(\alpha^*_{\text{emp}}, d_*)\} > 1 - n\eta$$

is valid.

§8 Selection of a Sample for Estimation of Values of an Arbitrary Function

In Chapter 8 we have shown that when a nonindicator function is estimated, the selection of a training sequence may lead to a function with a smaller guaranteed value of the expected risk. In the method of minimizing the overall risk the selection of a complete sample may result in a yet more significant effect. For a function linear in its parameters this additional effect arises because the exclusion of some vectors from the complete sample $x_1, \ldots, x_{l+k}$ changes the geometry of location of vectors. This allows us to carry out a more meaningful ordering of the class of functions $F(x, \alpha)$.

Thus let a complete sample

$$x_1, \ldots, x_{l+k} \tag{10.50}$$

be given. Consider H^t_{l+k} different subsets $X_1, \ldots, X_{H^t_{l+k}}$, each of which is obtained by omitting from (10.50) at most t elements. Below we shall assume that for all subsets the condition (10.16) is fulfilled.

Now let a training sequence

$$x_1, y_1; \ldots; x_l, y_l \tag{10.51}$$

and a working sequence

$$x_{l+1}, \ldots, x_{l+k} \tag{10.52}$$

be defined on the initial set (10.50). The samples (10.51) and (10.52) induce on each of the subsets its own training and working samples, respectively.

Consider H^t_{l+k} problems of estimating values of a function at given points. For each problem r ($r = 1, 2, \ldots, H^t_{l+k}$) we shall define—using the method described above—its own structure on the class of linear functions:

$$S^r_1 \subset \cdots \subset S^r_n.$$

We then obtain that with probability $1 - \eta$, for each of the problems (separately), the bound

$$I^{(r)}_{\Sigma}(\alpha_{\text{emp}}) < \left[\frac{1 + \tau a(p) \dfrac{l - l^r_n}{l + k - t_r} \varkappa^r_*}{1 - \tau a(p) \dfrac{k - k^r_{\text{ex}}}{l + k - t_r} \varkappa^2_*} \right]_{\infty} I^{(r)}_{\text{emp}}(\alpha_{\text{emp}})$$

is valid, where $F(x, \alpha_{\text{emp}}) \subset S^r_d$ is a function which minimizes the empirical risk on the training sequence for this problem (index (r) indicates that the overall and empirical risks are computed over the elements belonging to the subset $X^{(r)}$) and $\varkappa^r_*$ is the smallest solution of the inequality

$$d\left(\ln \frac{l + k - t_r}{d} + 1 \right) + \ln \Gamma_{l - l^r_{\text{ex}}, k - k^r_{\text{ex}}}(\varkappa^r) \leq \ln \frac{\eta}{1.5}.$$

Here we use the following notation: $l - l^r_{\text{ex}}$ is the length of the training sequence in problem r; $k - k^r_{\text{ex}}$ is the length of the working sample in the problem $l^r_{\text{ex}} + k^r_{\text{ex}} = t_r$. With probability $1 - \eta$, simultaneously for S_d elements of all H^t_{l+k} problems, the inequalities

$$I^{(r)}_{\Sigma}(\alpha) < \left[\frac{1 + \tau a(p) \dfrac{l - l^r_{\text{ex}}}{l + k - t_r} \varkappa^r_*}{1 - \tau a(p) \dfrac{k - k^r_{\text{ex}}}{l + k - t_r} \varkappa^r_*} \right]_{\infty} I^{(r)}_{\text{emp}}(\alpha)$$

are fulfilled, where (unlike the preceding case) $\varkappa^r_*$ are the smallest solutions of the inequalities

$$d\left(\ln \frac{l + k - t_r}{d} + 1 \right) + \ln \Gamma_{l - l^r_{\text{ex}}, k - k^r_{\text{ex}}}(\varkappa) + \ln H^t_{l+k} \leq \ln \frac{\eta}{1.5}.$$

We now choose a problem for which the bound on the value of the overall risk is minimal.

Finally, enumerating over d and t (in practice $t < 5$–10), we obtain the best solution.

§9 Estimation of Values of an Indicator Function in the Class of Piecewise Linear Decision Rules

Consider now the second approach to defining equivalence classes. Let a finite set of vectors X constituting a complete sample be subdivided into d subsets

$$X_1, \ldots, X_d, \qquad \bigcup X_i = X, \qquad X_i \bigcap X_j = 0 \qquad (i \neq j).$$

Assign to this subdivision a set of piecewise linear decision rules where on each subset X_i a linear decision rule is defined. Thus consider a parametric family of decision rules

$$\theta(L(x, \alpha_1), \ldots, L(x, \alpha_d)) = \begin{cases} L(x, \alpha_1) & \text{for } x \in X_1, \\ \vdots & \\ L(x, \alpha_d) & \text{for } x \in X_d, \end{cases} \tag{10.53}$$

where $L(x, \alpha_r)$ is a linear decision rule ($L(x, \alpha_r) = \theta(\sum_{i=1}^{n} \alpha_i^r \varphi_i(x))$).

The capacity of such a piecewise linear class of decision rules equals

$$h = nd.$$

Observe that the definition of a class of decision rules (10.53) is determined by a method of subdividing the complete sample X into d subsets. In order to determine the required subdivision we shall study the geometry of this set from the point of view of its *taxonomic structure* (cf. Appendix to this chapter). For this purpose we construct a tree (Figure 21). On the lowest

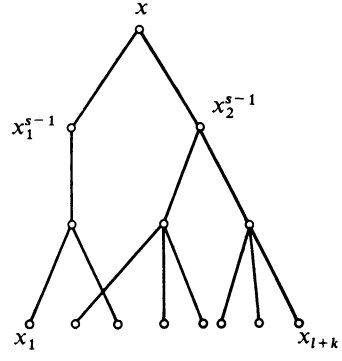

Figure 21

(sth) level of the tree each element of the set X forms a subset x_i. Elements of the $s-1$th level are formed by combining the elements of the sth level. Each element of the sth level is included in only one element of the $s-1$th level.

The highest (first) level consists of one subset X, which combines the whole set of elements of the population. The tree is constructed in such a way that at each level p the relation

$$\bigcup_{i=1} X_i^d = X, \quad X_i^d \bigcap X_j^d = 0 \ (i \neq j), \qquad d = 1, 2, \ldots, s \qquad (ns < l + k) \tag{10.54}$$

is valid. We assign to each level of the tree a family of piecewise linear decision rules S_d constructed in accordance with (10.54).

Using this method of constructing classes of piecewise linear decision rules, the taxonomic structure of the complete sample (10.54) will determine a specific structure on the class of piecewise linear decision rules:

$$S_1 \subset \cdots \subset S_s.$$

On this structure the method of structural minimization of the overall risk can be implemented, i.e., an element S_d of the structure can be found for which the method of minimizing the empirical risk will guarantee the smallest bound on the overall risk:

$$\nu_\Sigma(\alpha_{\text{emp}}) < \nu(\alpha^*_{\text{emp}}) + \frac{k\varkappa_*^2}{2(l+k)} + \varkappa_* \sqrt{\nu(\alpha^*_{\text{emp}}) + \left[\frac{k\varkappa_*}{2(l+k)}\right]^2} = R^*,$$

where $\varkappa_*$ is the smallest solution of the inequality

$$np\left(\ln \frac{l+k}{p} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leq \ln \frac{\eta}{1.5}.$$

The elements of the working sample are then classified using the obtained decision rule $F(x, \alpha^*_{\text{emp}})$. For such a classification the inequality

$$P\{\nu_\Sigma(\alpha^*_{\text{emp}}) < R^*\} \leq 1 - s\eta$$

is valid.

Methods of constructing taxonomic structures are considered in the Appendix to this chapter.

§10 Estimation of Values of an Arbitrary Function in a Class of Piecewise Linear Functions

A structure analogous to one considered in the preceding section can also be defined on a set of piecewise linear functions. For this purpose we shall consider the same taxonomic structure of the set $x_1, \ldots, x_{l+k}$ and determine

an element of the structure (i.e., a subdivision of the set $x_1, \ldots, x_{l+k}$ into taxons) for which the solution of the problem of minimizing the empirical risk in the class of linear decision rules separately for each taxon will assure the minimal guaranteed value of the overall risk.

For implementation of the method of structural minimization of the overall risk in the problem of estimating dependences the very same idea is used: for each element of the given taxonic structure S_d we obtain the minimal guaranteed bound on the value of the overall risk:

$$I_\Sigma(\alpha_{\text{emp}}) < \left[\frac{1 + \tau a(p)\dfrac{l}{l+k}\varkappa_*}{1 - \tau a(p)\dfrac{k}{l+k}\varkappa_*}\right] I_{\text{emp}}(\alpha_{\text{emp}}),$$

where $F(x, \alpha_{\text{emp}})$ is the function which minimizes the empirical risk in S_d, and $\varkappa_*$ is the smallest solution of the inequality

$$nd\left(\ln\frac{l+k}{nd} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leq \ln\frac{\eta}{1.5}$$

(here n is the dimension of the space, $nd < l + k$). We then select an element of the structure S_* and select on it the function $F(x, \alpha^*_{\text{emp}})$ for which the minimal guaranteed bound on the value of the overall risk for a given structure is obtained. The values of the function $F(x, \alpha^*_{\text{emp}})$ at the points of the working sample will be the estimated values of the function.

§11 Local Algorithms for Estimating Values of Indicator Functions

Finally let us consider the third approach for constructing an algorithm for estimating functional values. We define for each vector x^* of the complete sample a system of neighborhoods:

$$(x^*)_1, (x^*, x_{i_1})_2, \ldots, (x^*, x_{i_1}, x_{i_r})_r, \ldots, (x_1, \ldots, x_{l+k})_q.$$

Thus $l + k$ systems of neighborhoods are defined, a system for each vector of the complete sample:

$$\begin{array}{ll} (1) & (x_1)_1 \in (x_1, x_{i_1})_2 \in \cdots \in (x_1, \ldots, x_{l+k})_q; \\ (2) & (x_2)_1 \in (x_2, x_{i_2}, x_{i_3})_2 \in \cdots \in (x_1, \ldots, x_{l+k})_q; \\ \vdots & \\ (l+k) & (x_{l+k})_1 \in (x_{l+k}, x_{i_{l+k}})_2 \in \cdots \in (x_1, \ldots, x_{l+k})_q. \end{array} \tag{10.55}$$

Now let a subdivision of the set X into training and working samples be carried out.

Consider an arbitrary neighborhood X_i^r of the point x_i containing elements of both the training and working samples. In view of Theorem 10.2 one can assert with probability $1 - \eta$ that simultaneously for all linear decision rules the inequality

$$v_{\kappa}^{(r)}(\alpha) < \nu^{(r)}(\alpha) + \frac{k_r(\varkappa_*^{(r)})^2}{2(l_r + k_r)} + \varkappa_*^{(r)} \sqrt{\nu^r(\alpha) + \left[\frac{k_r \varkappa_*^{(r)}}{2(l_r + k_r)}\right]^2}$$

is fulfilled, where $v_{\Sigma}^{(r)}(\alpha)$ is the value of the overall risk of classification of elements belonging to the neighborhood X_i^r by means of a decision rule $F(x, \alpha)$, $\nu^{(r)}(\alpha)$ is the value of the empirical risk computed for the rule $F(x, \alpha)$ based on the elements of the training sequence belonging to the neighborhood X_i^r, $\varkappa_*^r$ is the smallest solution of the inequality

$$n\left(\ln \frac{l_r + k_r}{n} + 1\right) + \ln \Gamma_{l_r, k_r}(\varkappa) \leq \ln \frac{\eta}{1.5},$$

and n is the dimension of the space X. In this inequality l_r and k_r are the numbers of elements belonging to the neighborhood X_i^r in the training and working samples. Let $F(x, \alpha_{\text{emp}})$ be a decision rule which minimizes the value of the empirical risk on the training sequence belonging to X_i^r.

For the elements belonging to $X_i^{(r)}$ the bound

$$v_{\Sigma}^{(r)}(\alpha_{\text{emp}}) < \nu^{(r)}(\alpha_{\text{emp}}) + \frac{k_r(\varkappa_*^{(r)})^2}{2(l_r + k_r)} + \varkappa_*^{(r)} \sqrt{\nu^{(r)}(\alpha_{\text{emp}}) + \left[\frac{k_r \varkappa_*^{(r)}}{2(l_r + k_r)}\right]^2} = R_i(r)$$

is valid with probability $1 - \eta$. We shall now obtain a neighborhood of the point x_i for which the minimum (with respect to r) of the value $R_i(r)$ is attained. Let the minimum be attained in a neighborhood X_i^r, and let $\omega_{i_1}, \ldots, \omega_{i_s}$ be the classification of vectors obtained in the working sample belonging to this neighborhood. Clearly with probability $1 - \eta q$ this classification contains less than $R_i(\tau)k_\tau = R_i$ errors.

Analogously, solutions can be obtained for neighborhoods of all vectors belonging to the population. The results are presented in Table 1. In the

Table 1

Neighborhood of point	Classification of vectors					Bound on value of overall risk
	x_{l+1}	$\cdots$	x_{l+j}	$\cdots$	x_{l+k}	
x_1	ω_1'	$\cdots$	—	$\cdots$	ω_{l+k}'	R_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_s	—	$\cdots$	ω_{l+j}^s	$\cdots$	—	R_s
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_{l+k}	—	$\cdots$	—	$\cdots$	ω_{l+k}^{l+k}	R_{l+k}

first column of the table vectors are given which define the system of neighborhoods, followed by the best classification of vectors for the given system and finally the guaranteed bound on the number of classification errors. Observe that the same vectors of the working sample belong to the neighborhoods of different vectors and that the classifications of some vectors from the working sample presented in different rows of the second column may not be the same.

Denote by $\omega_1^*, \ldots, \omega_{l+k}^*$ the correct classification of vectors from the working sample $x_{l+1}, \ldots, x_{l+k}$. Then the content of the table may be written in the form

$$\begin{aligned} &\sum_{i}^{(1)} |\omega_{l+i}^* - \omega_{l+i}| < R_1, \\ &\qquad\qquad \vdots \\ &\sum_{i}^{(l+k)} |\omega_{l+i}^* - \omega_{l+i}| < R_{l+k}. \end{aligned} \tag{10.56}$$

Here $\sum^{(r)}$ indicates that the summation is carried out only over those classifications of vectors of the working sample which belong to the selected neighborhood of the point x_r.

Each one of the inequalities (10.56) is fulfilled with probability $1 - q\eta$. Consequently the system is consistent (all the inequalities are fulfilled simultaneously) with probability exceeding $1 - q(l + k)\eta$.

Consider the set Ω of vectors $\bar{w} = (\bar{\omega}_{l+1}, \ldots, \bar{\omega}_{l+k})$ of solutions of the system of inequalities (8.56). Actually the final vector of the classification may be chosen arbitrarily from this set. However, it is more expedient in such cases to choose a solution which possesses some additional extremal properties.

Among all the vectors in Ω we shall find the minimax one, w_m, i.e., the one whose distance from the farthest vector belonging to the admissible set Ω is the smallest:

$$w_m = \arg \min_{w \in \Omega} \max_{\bar{w} \in \Omega} |\bar{w} - w|.$$

The vector w_m will be chosen as the final solution of the problem of classifying vectors in the working sample.

In this algorithm, by defining a system of neighborhoods of vectors in the complete sample we were able to determine for each vector x_i an optimal neighborhood for constructing a linear decision rule. The rule thus obtained was used only for classification of vectors belonging to an optimal neighborhood. Such algorithms are sometimes referred to as *local* ones.

In practice different ideas for defining neighborhoods are utilized. In particular a neighborhood X_i^r of the vector x_i can be defined by means of metric closeness. (The set X_i^r contains vectors belonging to the complete sample such that $\|x_i - x\| \leq c$, where c is a constant. The collection of constants $c_1 < \cdots < c_t$ determines the system of neighborhoods.)

§12 Local Algorithms for Estimating Values of an Arbitrary Function

Using the scheme described in the preceding section one can immediately construct local algorithms for estimating values of a function of an arbitrary nature. Form a system of neighborhoods for vectors belonging to a complete sample:

$$\begin{aligned}
&(1) \qquad (x_1)_1 \in (x_1, x_{i_1})_2 \in \cdots \in (x_1, \ldots, x_{l+k})_q,\\
&\vdots\\
&(l+k) \quad (x_{l+k})_1 \in (x_{l+k}, x_{i_{l+k}})_2 \in \cdots \in (x_1, \ldots, x_{l+k})_q.
\end{aligned}$$

Let a subdivision of the set of vectors from the complete samples into elements belonging to training and working samples be carried out. Consider a system of neighborhoods for the point x_i:

$$X_i^1 \subset X_i^2 \subset \cdots \subset X_i^q,$$

$$X_i^r = (x_i, x_{i_2}, \ldots, x_{i_p})_r.$$

For each set X_i^r one can determine—using algorithms for estimating a linear function—the values of the function as well as a guaranteed bound on the value of the overall risk:

$$I_{\Sigma}^{(r)}(\alpha_{\text{emp}}) < \left[\frac{1 + \tau a(p)\dfrac{l_r}{l_r + k_r}\varkappa_*}{1 - \tau a(p)\dfrac{k_r}{l_r + k_r}\varkappa_*}\right]_{\infty} I_{\text{emp}}^{(r)}(\alpha_{\text{emp}}), \tag{10.57}$$

where $\varkappa_*$ is the smallest solution of the inequality

$$n\left(\ln\frac{l_r + k_r}{n} + 1\right) + \ln \Gamma_{l_r, k_r}(\varkappa) \leq \ln\frac{\eta}{1.5}. \tag{10.58}$$

Here l_r and k_r are the number's of elements in the training sequence and the working sample belonging to X_i^r.

Choose a neighborhood of point x_i and a function $F(x, \alpha_m^*)$ for which the bound (10.57) is minimal. Let k_r^* be the number of elements of the working sample belonging to this neighborhood. The inequality

$$\frac{1}{k_r^*}\sum_{k_r^*}(y_j - F(x_j, a_{\text{emp}}^*))^2 < \varkappa_r \tag{10.59}$$

is valid with probability $1 - q\eta$ for the values y_j belonging to this neighborhood obtained using the function $F(x, \alpha_{\text{emp}}^*)$. In (10.59) the summation is carried out over the vectors x from the working sample which belongs to the optimal neighborhood; y are the actual (unknown to us) values of the functional dependence at the points of the working sample, and $F(x_i; \alpha_{\text{emp}}^*)$ are the computed values. Thus for each point x_i (there are $l + k$ such points

in toto, which is the number of vectors in the complete sample) the inequality (10.59) is valid with probability $1 - \eta$. Therefore with probability $1 - (l + k)q\eta$ all the $l + k$ inequalities

$$\begin{aligned} \frac{1}{k_1^*} \sum_{k_1^*} (y_j - F(x_j, \alpha^*_{\text{emp}}(1)))^2 &< \varkappa_1, \\ &\vdots \\ \frac{1}{k^*_{l+k}} \sum_{k^*_{l+k}} (y_j - F(x_j, \alpha^*_{\text{emp}}(N)))^2 &< \varkappa_{l+k} \end{aligned} \tag{10.60}$$

are fulfilled simultaneously.

Consider an admissible set $\{Y\}$ of solutions $(y_{l+1}, \ldots, y_{l+k})$ of the system (10.60). This set is nonvoid with probability $1 - q(l + k)\eta$. Choose as the response a solution Y^* such that its distance from the farthest point in $\{Y\}$ is the smallest (a minimax solution), i.e., a k-dimensional vector Y^* for which the equality

$$Y^* = \arg \min_{Y \in \{Y\}} \max_{\bar{Y} \in \{Y\}} \rho(Y, \bar{Y})$$

is valid.

§13 The Problem of Finding the Best Point of a Given Set†

When the number of empirical data

$$x_1, y_1; \ldots; x_l, y_l \tag{10.61}$$

is small the statements of new problems are inspired by the idea that the intermediate problem may be more involved than the desired one.

It has been noted that the unknown density need not be estimated in order to estimate the function. There is no point in estimating a function if all we want to know is its values at given points

$$x_{l+1}, \ldots, x_{l+k} \tag{10.62}$$

In this section we proceed from the assumption that we should not generally estimate the values of functions at the points (10.62) if our goal is to find *the best point in the set* (10.62) i.e., the point of which one can assert with the highest probability that the function unknown to us which specifies the value of y takes on its highest (or lowest) value there. As in earlier similar situations, a case is possible where the available data (10.61) and (10.62) are insufficient for satisfactory solution of the intermediate problem (estimating the values of the function in all the points of the set (10.62)), but are sufficient to solve the desired problem (to find the best point of the set (10.62)).

† Section 13 was translated by the author.

Below we shall specify for a certain situation a technique whereby the best point of the set is found; but first we should like to note that statement of this problem is a response to the limited amount of available empirical data in the solution of important real-life problems.

EXAMPLE. Only a few dozen antitumor drugs have been clinically tested in the world by now. Meanwhile hundreds of new antitumor drugs are synthesized annually. These are tested in different models of human tumors (including animal tumors). Effectiveness in models does not, however, insure its clinical effectiveness. The problem is to be able to identify the clinically most active drugs among the newly synthesized drugs using the results of model tests with these drugs, the information of clinical activity of drugs that have been already tested in clinics, and the information of the activity of the same drugs in various models [80a].

Thus, let a learning sample (10.61) and a working sample (10.62) be given. Let a class of functions $F(x, \alpha)$ be so specified that it contains a function $F(x, \alpha_0)$ that orders the vectors x of the learning and working samples in decreasing order of values of $F(x, \alpha_0)$ in the same way as an unknown function $f(x)$ which determines the values of y. (For indicator functions this condition degenerates into the requirement that $f(x)$ belong to the class of $F(x, \alpha)$.) Among vectors of the working sample it is required to find a vector x_* of which one can assert with the highest probability that the function $f(x)$ takes on the largest value on it.

As before, we shall isolate the case where $F(x, \alpha)$ is an indicator function, or $y = \omega$. In this case it is required that among vectors of the working sample a vector x_* should be indicated for which the probability of classification $\omega_* = 1$ is maximal.

Note that for this particular case the problem of choosing the best point of a set becomes degenerate. Under the conditions where $y = \omega$ can be either zero or one, finding the best point is generally equivalent to indicating its value. In the general case where y takes on an arbitrary (>2) number of values it is required to indicate the best point (rather than its values).

13.1 Choice of the Most Probable Representative of a Given Class

Let us first consider a case where $y = \omega$ and $F(x, \alpha)$ is the class of indicator functions. We shall denote

$$\begin{aligned} R &= x_1, y_1; \ldots; x_l, y_l \\ X &= x_{l+1}; \ldots; x_{l+k}, \\ X_i &= x_{l+1}, \ldots, \hat{x}_i, \ldots, x_{l+k}. \end{aligned} \tag{10.63}$$

(The sequence X_i is obtained from X by omitting the element x_i.) The sequence X_i can be divided into two classes in 2^{k-1} possible ways. Let

$$\Omega_r^i = \omega_{l+1}^r, \ldots, \hat{\omega}_i, \ldots, \omega_{l+k}, \qquad r = 1, 2, \ldots, 2^{k-1}$$

denote the rth way. Assume that for each $r = 1, 2, \ldots, 2^{k-1}$ the probability $P(\Omega_r^i)$ that Ω_r^i will coincide with the classification of the sequence X_i performed with the aid

of the function $f(x)$ has been defined. Then for each fixed vector x_i of the working sample,

$$P\{\omega_i = 1 | R, X\} = \sum_{r=1}^{2^{k-1}} P\{\omega_i = 1 | R, X, \Omega_r^i\} P(\Omega_r^i). \tag{10.64}$$

Moreover, since the class $F(x, \alpha)$ contains the function $f(x)$, then faultless division of the complete set of vectors is possible with the use of one of the N equivalence classes, $F_1, \ldots, F_N$. Let us assume *a priori* that the faultless division of vectors can be equiprobably performed by any equivalence class.

The probabilities $P(\Omega_r^i)$ and $P\{\omega_i = 1 | R, X, \Omega_r^i\}$ on the right-hand side of the equality (10.64) can be then immediately calculated. Namely, the probability that the classification Ω_r^i of the vectors X_i coincides with that specified by the function $f(x)$ is equal to

$$P(\Omega_r^i) = \frac{n(X_i, \Omega_r^i)}{N}, \tag{10.65}$$

where $n(X_i, \Omega_r^i)$ is the number of equivalence classes which classify the sequence X_i in compliance with Ω_r^i. The conditional probability that the vector x_i belongs to the class $\omega_i = 1$ is equal to

$$P\{\omega_i = 1 | R, X, \Omega_r^i\} = \chi_r^i = \begin{cases} 0 & \text{if there is no equivalence class which permits the division} \\ & R \cup X_i \Omega_r^i; \\ \frac{1}{2} & \text{if there is an equivalence class which permits the division} \\ & R \cup X_i \Omega_r^i \cup x_i, 1 \\ & \text{and a class which permits the division} \\ & R \cup X_i \Omega_r^i \cup x_i, 0; \\ 1 & \text{if there is an equivalence class which permits the division} \\ & R \cup X_i \Omega_r^i \cup x_i, 1, \\ & \text{and there is no equivalence class which permits the division} \\ & R \cup X_i \Omega_r^i \cup x_i, 0. \end{cases} \tag{10.66}$$

Substituting the expressions (10.65) and (10.66) into (10.64), we have that the probability that the vector x_i belongs to the class $\omega_i = 1$ is equal to

$$P\{\omega_i = 1 | R, X\} = \sum_{r=1}^{2^{k-1}} \frac{\chi_r^i n(X_i, \Omega_r^i)}{N}. \tag{10.67}$$

What remains to do is to choose from k vectors of the sample the vector for which this probability is maximal.

Note that the wider the class of functions $F(x, \alpha)$, the smaller generally is

$$\max_i P\{\omega_i = 1 | R, X\}.$$

In the limiting case where the class of functions $F(x, \alpha)$ is so broad that it permits the maximal possible number $N = 2^{l+k}$ of equivalence classes, the equality

$$P\{\omega_i = 1 | R, X\} = \tfrac{1}{2}$$

holds no matter what the number i is.

Another natural assumption is that the *a priori* probability of faultless division of the complete sample by the rules from F_j is given by the binomial law with a parameter p (p is the probability that an element with $\omega = 1$ will occur). The *a priori* probability of faultless division is

$$p_j = \frac{C_{l+k}^{m_j} p^{m_j} (1-p)^{l+k-m_j}}{\sum_{j=1}^{N} C_{l+k}^{m_j} p^{m_j} (1-p)^{l+k-m_j}},$$

where m_j is the number of elements which are classified by the rules F_j with $\omega = 1$. Here, in place of (10.67) we have

$$P(\omega_i = 1 | R, X) = \sum_{r=1}^{2^{k-1}} \hat{\chi}_r^i \frac{\sum_{j=1}^{n(X_i, \Omega_r^i)} C_{l+k}^{m_j} p^{m_j} (1-p)^{l+k-m_j}}{\sum_{j=1}^{N} C_{l+k}^{m_j} p^{m_j} (1-p)^{l+k-m_j}},$$

where

$$\hat{\chi}_r^i = \begin{cases} 0 & \text{if there is no equivalence class which permits the division} \\ & R \cup X_i \Omega_r^i; \\ \dfrac{p}{p + \dfrac{m_*^r (1-p)}{l+k+1-m_*^r}} & \text{if there is an equivalence class which permits the division} \\ & R \cup X_i \Omega_r^i \cup x_i, 1 \\ & \text{and class which permits the divisions} \\ & R \cup X_i \Omega_r^i \cup x_i, 0; \\ 1 & \text{if there is an equivalence class which permits the division} \\ & R \cup X_i \Omega_r^i \cup x_i, 1 \\ & \text{and there is no equivalence class which permits the division} \\ & R \cup X_i \Omega_r^i \cup x_i, 0, \end{cases}$$

and m_*^r is the number of pairs in the set $R \cup X_i \Omega_r^i \cup x_i, 1$ with $\omega = 1$.

13.2 Choice of the Best Point of a Given Set

Consider a general case where in the learning sequence

$$x_1, y_1; \ldots; x_l, y_l,$$

y can take on arbitrary values. Note that elements of the sequence

$$X = x_{l+1}, \ldots, x_{l+k}$$

can be ordered in all possible ways by using permutation operators $T_r (r = 1, 2, \ldots, k!)$.

Let $T_0 X_0$ denote a sequence which consists of vectors x of the learning sequence and is ordered in decreasing corresponding values of y (for simplicity let us assume that the ordering is strict).

Let us write

$$z \overset{f}{\succ} X$$

if $f(z) > f(x)$ for all $x \in X$, i.e., if z precedes all elements of the set with ordering in terms of values of $f(x)$.

Assume now that on the working sample a vector x_i has been fixed. There are $(k-1)!$ different ways to order the set X_i. Assume that for each of these the probability $P\{T_r X_i\}$ has been determined that the ordering $T_r X_i$ will coincide with the ordering of the vectors x of the set X_i in decreasing value of the function $f(x)$. Then for each fixed vector x_i of the working sample,

$$P\{x_i \overset{f}{\succ} X_i | T_0 X_0, X\} = \sum_{r=1}^{(k-1)!} P\{x_i \overset{f}{\succ} X_i | T_0 X_0, T_r X_i\} P\{T_r X_i\}. \tag{10.68}$$

From the viewpoint of ordering the vectors x of the complete sample, the class of functions $F(x, \alpha)$ decomposes into a finite number of equivalence classes $F_1, \ldots, F_N$ (each containing functions which order the vectors x of the complete set in the same way).

In compliance with the conditions of the problem, among all equivalence classes there is one which orders complete sample vectors as does the function $f(x)$ which specified the values of y. Let us assume *a priori* that any of the N equivalence classes can be this class with the same probability. Then, as in the particular case, the probabilities $P\{T_r X_i\}$ and $P\{x_i \overset{f}{\succ} X_i | T_0 X_0, T_r X_i\}$ can be computed. The probability that the ordering $T_r X_i$ will coincide with the ordering in decreasing values of the function $f(x)$ is

$$P\{T_r X_i\} = \frac{n(T_r X_i)}{N} \tag{10.69}$$

where $n(T_r X_i)$ is the number of equivalence classes which permit ordering $T_r X_i$. The conditional probability that x_i is the best point is

$$P\{x_i \overset{f}{\succ} X_i | T_0 X_0, T_r X_i\} = \hat{\chi}_r^i = \left\{ \begin{array}{c} 0 \\ \vdots \\ \dfrac{1}{p+1} \\ \vdots \\ 1 \end{array} \right\}, \tag{10.70}$$

where χ_r^i is equal to zero if there is no equivalence class which permits simultaneous ordering to $T_0 X_0$ and $x_i T_r X_i$ ($x_i T_r X_i$ is a sequence whose leftmost element is x_i and the

remaining ones coincide with $T_r X_i$); and $\hat{\chi}_r^i$ is equal to $1/(p+1)$ $(p = 0, 1, \ldots, k-1)$ if there is an equivalence class which permits simultaneous ordering of $T_0 X_0$ and $x_i T_r X_i$ and there are p other equivalence classes which permit simultaneous ordering of $T_0 X_0$ and $x_i T_r X_i$

Substituting (10.69) and (10.70) into (10.68), we have

$$P\{x_i \overset{f}{\succ} X_i | T_0 X_0, X\} = \sum_{r=1}^{(k-1)!} \frac{\hat{\chi}_r^i n(T_r X_i)}{N}. \tag{10.71}$$

Consequently, the probability that among vectors of the working sample the vector x_i will have the maximal value of $f(x_i)$ is determined by the expression (10.71). The vector for which the probability is the largest should be chosen.

As before, the broader the class of the functions $F(x, \alpha)$, the smaller in general is

$$\max_i P\{x_i \overset{f}{\succ} X_i | T_0 X_0, X\}.$$

In the limiting case where the class of $F(x, \alpha)$ is so broad that the maximal possible number $N = (l+k)!$ of equivalence classes is permitted, the equality

$$P\{x_i \overset{f}{\succ} X_i | T_0 X_0, X\} = \frac{1}{k}$$

holds irrespective of the number of i.

§14 Remarks on Estimating Values of a Function

Remark 1. We have seen that the construction of methods for estimating values of a function at given points is related to various methods of allocating $l + k$ vectors in an n-dimensional space. In this chapter we have studied the geometry of $l + k$ vectors in a complete sample from three different points of view. We have investigated:

(1) the structure of linear subdivisions of vectors in a complete sample,
(2) a taxonomic structure of vectors in a complete sample,
(3) the structure of neighborhoods of elements in a complete sample.

A description of the structure of a complete sample from each one of these points of view generates its own method of estimating values of a function at given points.

The methods of studying the geometry of a complete sample presented above do not exhaust all the possible procedures for examining the mutual arrangement of $l + k$ vectors in an n-dimensional space. Other methods are available, and each one may serve as a basis for devising a method of estimating functional values.

Remark 2. In this chapter, when estimating values of functions at given points, we have assumed that the risk is determined by a quadratic loss function

$$(y - F(x, \alpha))^2.$$

However, all the results obtained herein can be extended to the case of a more general loss function

$$\Phi(y - F(x, \alpha)).$$

Remark 3. The smaller the size of the training sample, the greater the effect due to direct estimation of values of a function at given points as compared with the traditional methods: estimating a function by means of a training sequence and computing its values. This effect is illustrated in Table 2, based on data which were available for the solution of problems of medical differential diagnostics using the method of pattern recognition. The first column of the table gives the identification number of the experiment; the second, the size of the training sequence; the third, the size of the working sample; the fourth, the number of classification errors on the working

Table 2

No.	l	l_p	m	$\hat{m}$
1	12	21	6	3
2	24	21	5	2
3	23	10	3	1
4	27	14	6	3
5	29	28	9	3
6	49	35	13	9
7	42	35	10	6
8	52	35	12	8
9	65	46	14	8
10	33	57	18	5

sample using a linear decision rule which minimizes the empirical risk (the method of a generalized portrait; cf. Addenum 1); the fifth, the number of classification errors incurred in classifying the elements of the working sample using the method of minimizing the overall risk. In these problems the initial dimension of the space of binary features was 60. The problems were solved using algorithms presented in Addendum I.

Appendix to Chapter 10

Taxonomy Problems

§A1 A Problem of Classification of Objects

Let it be required to subdivide the set of objects

$$X = x_1, \ldots, x_l \tag{A.1}$$

into subsets

$$X_1, \ldots, X_m \tag{A.2}$$

such that the following two conditions are fulfilled:

(1) the subsets are disjoint, i.e.,

$$X_i \cap X_j = 0 \qquad (i \neq j); \tag{A.3}$$

(2) any element belonging to (A.1) falls into one of the subsets (A.2), i.e.,

$$\bigcup_{i=1}^{m} X_i = X, \tag{A.4}$$

and each one of the subsets should consist of "the most similar elements."

In other words, it is required under the conditions (A.3) and (A.4) to minimize a functional which is defined on the set of all subdivisions of the set X and which reflects the notion of the quality of a subdivision of the set X.

The subsets $X_1, \ldots, X_m$ which serve as an optimal solution of such a problem are called *taxons*, and the problem of subdividing a set X into subsets is referred to as a *problem of taxonomy*.

Thus the problem is to write down a functional which reflects our conception of the quality of subdivision of a set and to obtain a subdivision which yields a minimal value for this functional.

The problem of constructing the functional is an informal one—each investigator may determine his or her own concept of an optimal solution. Nevertheless there exists a "natural" definition of the quality of a solution for a particular problem of taxonomy, namely the problem of subdividing the initial set X into m taxons $X_1, X_2, \ldots, X_m$ (where the number m is given in advance). In this case a number $d(X_i)$ is determined which characterizes for each subset X_i the degree of closeness (or similarity) of its objects. By means of these quantities $d(X_i)$ a functional

$$d = \sum_{i=1}^{m} d(X_i) \tag{A.5}$$

is formed.

In the theory of taxonomy the following characterizations of closeness between objects in a set X_i are adopted:

(1) the mean squared deviation from the center of gravity of the sets:

$$d_1(X_i) = \frac{1}{l_i} \sum_{j=1}^{l_i} (x_{ij} - \bar{x}_i)^{\mathrm{T}}(x_{ij} - \bar{x}_i),$$

where l_i is the number of elements in the set X_i, $\bar{x}_i = (1/l_i) \sum_{j=1}^{l_i} x_{ij}$ is the center of gravity of the set X_i, and x_{ij} are the elements in the set X_i;

(2) the mean squared deviations between the elements of the set

$$d_2(X_i) = \frac{1}{l_i(l_i - 1)} \sum_{\substack{j,t \\ i>t}} (x_{ij} - x_{it})^{\mathrm{T}}(x_{ij} - x_{it});$$

(3) the value of the determinant of the dispersion (variance–covariance) matrix for the vectors of the set

$$|d_3(X_i)| = \left| \frac{1}{l} \sum_{j=1}^{l} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^{\mathrm{T}} \right|.$$

From these quantities the following functionals are formed:

$$\begin{aligned} d_1 &= \sum_{i=1}^{m} d_1(X_i), \\ d_2 &= \sum_{i=1}^{m} d_2(X_i), \\ d_3 &= \left| \sum_{i=1}^{m} d_3(X_i) \right|. \end{aligned} \tag{A.6}$$

Here $|\sum_{i=1}^{m} d_3(X_i)|$ is the determinant of the matrix $\sum_{i=1}^{m} d_3(X_i)$.

Other functionals are available which reflect different conceptions of quality for solutions of the particular taxonomy problem (when the number of taxons is given in advance). These functionals are presented in papers

[1, 2, 19]. As far as the general taxonomy problem is concerned (where the number of taxons is not known in advance), there are at present no widely acceptable definitions of the quality of taxonomy.

§A2 Algorithms of Taxonomy

The problem of minimizing the functional (A.5) on the set of possible subdivision of l objects into m groups is a problem of discrete programming: altogether there are

$$N(l, m) = \sum_{i=0}^{m} (-1)^i C_m^i (m - i)^l$$

different subdivisions of l objects into m groups in such a manner that no group will be empty. It is necessary to choose among the $N(l, m)$ subdivisions the one which minimizes the functional (A.5).

An exact solution of this problem requires a large number of calculations (of the order of magnitude of the value $N(l, m)$). Therefore for solving taxonomy problems heuristic methods are used; in particular, the following procedure: Two sequences $\tilde{X}_l = \tilde{x}_1, \ldots, \tilde{x}_l$, $Q = q_1, \ldots, q_l$ are constructed on the set of vectors $X(x_1, \ldots, x_l)$ by means of the following inductive rule:

(1) First an arbitrary element belonging to X is selected say x_1, and we set $\tilde{x}_1 = x_1$, $q_1 = 0$. The vector x_1 is then excluded from the set X, thus forming the set M_1 ($M_1 = X/x_1$).

(2) Then t vectors are chosen before the $(t + 1)$th step from the initial set, and the sequences

$$\tilde{X}_t = \tilde{x}_1, \ldots, \tilde{x}_t,$$

$$Q_t = q_1, \ldots, q_t$$

are constructed, while the remaining vectors are combined into the set M_t. Then at the $(t + 1)$st step the vector in M_t which is closest to the sequence $\tilde{X}_t$ is adjoined to that sequence, i.e., a vector $x = \tilde{x}_{t+1}$ for which the minimum

$$q_{t+1} = \min_{x_i \in M_t} \rho(x_i, \tilde{X}_t)$$

is attained. This vector is added to the constructed sequence, forming the sequence $\tilde{X}_{t+1}$, and the corresponding quantity q_{t+1} is added to Q_t, thus forming a new sequence Q_{t+1}. On the other hand the vector $\tilde{x}_{t+1}$ is excluded from the set M_t forming the set M_{t+1}. This process is continued until all the vectors belonging to X are ordered.

(3) Utilizing the sequence $Q_l = q_1, \ldots, q_l$, one can subdivide the sequence $\tilde{X}_l = \tilde{x}_1, \ldots, \tilde{x}_l$ into m taxons. For this purpose we choose a number

q^* such that only $m - 1$ values of the sequence $Q_l = q_1, \ldots, q_l$ exceed q^*. Let $q_{i_1}, q_{i_2}, \ldots, q_{i_{m-1}}$ be the corresponding values of the sequence Q_{l+1}.

Then the vectors belonging to the sequence $\tilde{X}_l$ with indices from 1 to i_1 form the first taxon; vectors with indices from i_1 to i_2 form the second one, and so on. In all there will be m taxons.

To construct a specific algorithm it is necessary to define the distance between a point x and a set X. Usually the following metric $\rho(x, X)$ is used: the distance from x to the closest vector belonging to X.

As far as the distance between two vectors x_a and x_b belonging to X is concerned, in addition to the usual Euclidean metric in taxonomy, certain special metrics are used. In particular, *Tanimoto's metric* is encountered, which defines the proximity between two sets X and Y as follows:

$$\rho_T(X, Y) = \frac{n_x + n_y - 2n_{xy}}{n_x + n_y - n_{xy}}.$$

here n_x is the number of elements in X, n_y is the number of elements in Y, and n_{xy} is the number of elements which appear simultaneously in both sets.

Using the Tanimoto metric, the proximity between the set of objects $X^\delta(x_a)$ belonging to X which fall into a δ-neighborhood of the point x_a and a set of objects $X^\delta(x_b)$ which fall into a δ-neighborhood of point x_b is determined. Thus

$$\rho(x_a, x_b) = \rho_T(X^\delta(x_a); X^\delta(x_b)).$$

(here δ is a given parameter).

Such a metric is more suitable for studying "the geometry of vectors as a whole."

Postscript

In this book the problem of estimating dependences from empirical data has been studied from the standpoint of approximating functions. Two new ideas were implemented:

(1) a definition of a structure on the class of functions in which the estimation is carried out and minimization of the risk over the elements of the structure (the method of structural risk minimization),
(2) partitioning into equivalence classes and a definition of the structure on these classes (the estimation of values of a function at given points).

It was shown that the development of these ideas results in devising more precise methods of estimation than the traditional ones.

However, all the specific structures studied in this book have arisen from commonsense considerations rather than as a result of explorations and analysis. Moreover the definition of a structure adopted in this book satisfies the axiomatization of algebraic structures. Therefore one may expect that using analytic methods it will be possible to find structures which are more meaningful and suggestive than those which are utilized in this book. There are no investigations at present in this direction.

Addendum I

Algorithms for Pattern Recognition

§1 Remarks about Algorithms

In the main part of the book the theory of estimating dependences from empirical data was presented. Classical methods of estimating dependences (Chapters 3, 4, and 5) were considered. These methods are efficient under the conditions that the required dependence belongs to a given class, and they guarantee the determination of a satisfactory solution when the training sample is sufficiently large.

In practice we cannot be sure that the required dependence belongs to a class of functions in which the estimation is carried out, nor can we be confident that the sample size at hand is sufficient to arrive at a good approximation. Therefore methods of minimizing risk were developed which do not require the knowledge of the model of the desired dependence and are geared towards utilization of samples of limited size (Chapter 6–10).

The addenda to this book are devoted to problems of constructing estimation algorithms.

In this addendum we shall consider algorithms for pattern recognition problems. The algorithms are based on the utilization of bounds on the uniform relative deviation of frequencies from their probabilities which are valid for any probability measure $P(x, \omega)$ (including the least favorable one).

Usually when it comes to the construction of algorithms based on a certain theory it turns out that the theory developed is only a rough approximation to reality. As a rule this roughness is compensated by the fact that in the course of constructing algorithms, theories are not followed verbatim. Authors of algorithms express their own understanding of reality which cannot be formalized. This is also true here.

In practice there is no reason to assume that the least favorable distribution $P(x, \omega)$ will be realized. Therefore estimates derived from a general theory in actual situations may be excessive. How then does one take into account that we are dealing with the actual distributions and not with the least favorable ones? The answer to this question determines the degree of informality in our approach to the constructed theory.

An informal approach to the theory when constructing algorithms for pattern recognition is reflected by the fact that along with the bound

$$P(\alpha) < \nu_{\text{emp}}(\alpha) + 2\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{12}}{l} \times\left(1+\sqrt{1+\frac{\nu_{\text{emp}}(\alpha)l}{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{12}}}\right) \tag{D-I.1}$$

when estimating indicator functions we shall assume that the bound

$$P(\alpha) < \nu_{\text{emp}}(\alpha) + \frac{h\left(\ln\frac{l}{h}+1\right)-\ln\eta}{2l}\left(1+\sqrt{1+\frac{4\nu_{\text{emp}}(\alpha)l}{h\left(\ln\frac{l}{h}+1\right)-\ln\eta}}\right) \tag{D-I.2}$$

(which differs from (D-I.1) in its constants) is valid. Moreover when estimating the values of indicator functions we shall assume that along with the bound

$$\nu_{\Sigma}(\alpha) < \nu_{\text{emp}}(\alpha) + \frac{k}{2(l+k)}\varkappa_*^2 + \varkappa_*\sqrt{\nu_{\text{emp}}(\alpha)+\left[\frac{k\varkappa_*}{2(l+k)}\right]^2}, \tag{D-I.3}$$

where $\varkappa_*$ is the smallest solution of the inequality

$$h\left(\ln\frac{l+k}{h}+1\right)+\ln\Gamma_{l,k}(\varkappa) \leqq \ln\frac{\eta}{1.5},$$

the bound

$$\nu_{\Sigma}(\alpha) < \nu_{\text{emp}}(\alpha) + \frac{k}{2(l+k)}\varkappa_*^2 + \varkappa_*\sqrt{\nu_{\text{emp}}(\alpha)+\left[\frac{k\varkappa_*}{2(l+k)}\right]^2} \tag{D-I.4}$$

is also valid, where $\varkappa_*$ is the smallest solution of the inequality

$$h\left(\ln\frac{l+k}{h}+1\right)+\ln\Gamma_{l,k}(\varkappa) \leqq \ln\eta. \tag{D-I.5}$$

This bound involves different constants than the bound (D-I.3).

§2 Construction of Subdividing Hyperplanes

Algorithms for constructing hyperplanes which subdivide two finite sets of vectors – the set of vectors

$$X_a : x_1, \ldots, x_a \tag{D-I.6}$$

and the set

$$\bar{X}_b : \bar{x}_1, \ldots, \bar{x}_b \tag{D-I.7}$$

—serve as the basis for constructing algorithms for pattern recognition in the class of linear decision rules. The problem reduces to finding a vector ψ for which the inequalities

$$\left.\begin{array}{ll} x_i^{\mathrm{T}}\psi \geqq 1 & \text{for } x_i \in X_a \\ \bar{x}_j^{\mathrm{T}}\psi \leqq k & \text{for } \bar{x}_j \in \bar{X}_b \end{array}\right\} \quad k < 1, \tag{D-I.8}$$

are fulfilled. Clearly if there exists a vector ψ for which the inequality (D-I.8) is fulfilled, we then have a set of vectors ψ satisfying (D-I.8). In this set we shall seek a vector which is the smallest in absolute value. This vector is called a *generalized portrait* [12].

Minimization of the quadratic form

$$\|\psi\|^2 = \psi^{\mathrm{T}}\psi \tag{D-I.9}$$

subject to the conditions (D-I.8) is a quadratic programming problem. Necessary and sufficient conditions for the minimum of (D-I.9) under the restrictions (D-I.8) are given by the Kühn–Tucker theorem.

Theorem D-I.1 (Kühn–Tucker). *Let a differentiable convex function $F(x)$ and linear functions $f_i(x)$ $(i = 1, 2, \ldots, l)$ be given. Let x_0 yield the minimum for $F(x)$ under the restrictions*

$$f_i(x) \geq 0. \tag{D-I.10}$$

Then there exist $\lambda_i \geq 0$ satisfying the conditions

$$\lambda_i f_i(x_0) = 0, \tag{D-I.11}$$

such that the equality

$$\nabla F(x_0) = \sum_{i=1}^{l} \lambda_i \nabla f_i(x_0) \tag{D-I.12}$$

is fulfilled (∇ is the gradient sign).

Conversely, if for some point x_0 the conditions (D-I.10) *are fulfilled and one can find $\lambda_i \geq 0$ satisfying the conditions* (D-I.11) *and* (D-I.12), *then at the point x_0 the conditional minimum for $F(x)$ under the restrictions* (D-I.10) *is attained.*

Proof of the Kühn–Tucker theorem are presented in all textbooks on convex programming (for example [65]).

We shall apply the Kühn–Tucker theorem to our case of minimizing (D-I.9) under the restrictions (D-I.8).

Theorem D-I.2. *The minimal in absolute-value vector ψ satisfying* (D-I.8) (*a generalized portrait*) *can be represented in the form*

$$\psi = \sum_{i=1}^{a} \alpha_i x_i - \sum_{j=1}^{b} \beta_j \bar{x}_j, \qquad \alpha_i \geq 0, \qquad \beta_j \geq 0, \tag{D-I.13}$$

where

$$\begin{aligned} \alpha_i[x_i^{\mathrm{T}}\psi - 1] &= 0, \qquad i = 1, 2, \ldots, a, \\ \beta_j[k - x_j^{\mathrm{T}}\psi] &= 0, \qquad j = 1, 2, \ldots, b. \end{aligned} \tag{D-I.14}$$

Among all the vectors ψ satisfying (D-I.8) *the vector ψ represented by* (D-I.13), (D-I.14) *is minimal in absolute value.*

The proof follows directly from the Kühn–Tucker theorem.

The vectors x_i, $\bar{x}_j$ for which the conditions

$$\begin{aligned} x_i^{\mathrm{T}}\psi &= 1, \qquad x_i \in X_a, \\ \bar{x}_j^{\mathrm{T}}\psi &= k, \qquad \bar{x}_j \in \bar{X}_b, \end{aligned} \tag{D-I.15}$$

are fulfilled will be called the *extreme vectors*. In view of Theorem D-1.2 a generalized portrait is decomposable with nonzero weights only in terms of a system of extreme vectors.

Consider now the *dual problem* whose solution is equivalent to constructing a generalized portrait. Introduce the space of parameters $E_{\alpha\beta}$, and consider the function

$$W(\alpha, \beta) = \sum_{i=1}^{a} \alpha_i - k \sum_{i=1}^{b} \beta_j - \tfrac{1}{2}\psi^{\mathrm{T}}\psi, \tag{D-I.16}$$

where the vector ψ is given by

$$\psi = \sum_{i=1}^{a} \alpha_i x_i - \sum_{j=1}^{b} \beta_j \bar{x}_j.$$

We shall show that the point α_0, β_0 which is a point of maximum for the function $W(\alpha, \beta)$ in the positive quadrant $\alpha_i \geq 0$, $\beta_j \geq 0$ determines a generalized portrait.

Indeed, necessary and sufficient conditions for the maximum of the function $W(\alpha, \beta)$ at the point α_0, β_0 are the conditions

$$\frac{\partial W(\alpha_0, \beta_0)}{\partial \alpha_i} \begin{cases} = 0 & \text{for } \alpha_i^0 > 0, \\ \leq 0 & \text{for } \alpha_i^0 = 0, \end{cases} \qquad i = 1, 2, \ldots, a,$$

$$\frac{\partial W(\alpha_0, \beta_0)}{\partial \beta_j} \begin{cases} = 0 & \text{for } \beta_j^0 > 0, \\ \leq 0 & \text{for } \beta_j^0 = 0, \end{cases} \qquad j = 1, 2, \ldots, b.$$

We shall write down these conditions using the notation

$$\psi_0 = \sum_{i=1}^{a} \alpha_i^0 x_i - \sum_{j=1}^{b} \beta_j^0 \bar{x}_j,$$

We thus obtain

$$\begin{aligned} 1 - x_i^T\psi_0 &\begin{cases} =0 & \text{for } a_i^0 > 0, \\ \leq 0 & \text{for } \alpha_i^0 = 0, \end{cases} \qquad i = 1, 2, \ldots, a, \\ \bar{x}_j^T\psi - k &\begin{cases} =0 & \text{for } \beta_j^0 > 0, \\ \leq 0 & \text{for } \beta_j^0 = 0, \end{cases} \qquad j = 1, \ldots, b. \end{aligned} \tag{D-I.17}$$

These conditions can be rewritten as inequalities

$$\begin{aligned} x_i^T\psi_0 &\geq 1, \qquad i = 1, 2, \ldots, a, \\ \bar{x}_j^T\psi_0 &\leq k, \qquad j = 1, 2, \ldots, b, \end{aligned} \tag{D-I.18}$$

and equalities

$$\begin{aligned} \alpha_i^0(1 - x_i^T\psi_0) &= 0, \qquad i = 1, 2, \ldots, a, \\ \beta_j^0(\bar{x}_j^T\psi_0 - k) &= 0, \qquad j = 1, 2, \ldots, b. \end{aligned}$$

In view of the assertion of Theorem D-I.2, these conditions determine a generalized portrait. Thus the problem of constructing a hyperplane subdividing two sets of vectors is reduced to the determination of a maximum for the function $W(\alpha, \beta)$ in the positive quadrant.

Below we shall consider methods for minimizing the quadratic form $W(\alpha, \beta)$ in the positive quadrant. First, however, we shall verify the following important fact.

Theorem D-I.3. *If the subdividing hyperplane exists (i.e., there exists a vector ψ_0 for which the inequalities* (D-I.18) *are fulfilled), then the maximum of the function $W(\alpha, \beta)$ in the positive quadrant equals one-half the square of the absolute value of the generalized portrait:*

$$W(\alpha_0, \beta_0) = \frac{\|\psi_0\|^2}{2}. \tag{D-I.19}$$

PROOF. According to Theorem D-I.2

$$\psi_0 = \sum_{i=1}^{a} \alpha_i^0 x_i - \sum_{j=1}^{b} \beta_j^0 \bar{x}_j,$$

Therefore

$$\|\psi_0\|^2 = \psi_0^T\psi_0 = \sum_{i=1}^{a} \alpha_i^0 x_i^T\psi - \sum_{j=1}^{b} \beta_j^0 \bar{x}_j^T\psi_0$$

and taking (D-I.15) into account, we obtain

$$\|\psi_0\|^2 = \sum_{i=1}^{a} \alpha_i^0 - k \sum_{j=1}^{b} \beta_j.$$

Thus,

$$W(\alpha_0, \beta_0) = \sum_{i=1}^{a} \alpha_i^0 - k \sum_{j=1}^{b} \beta_j^0 - \tfrac{1}{2}\psi_0^{\mathrm{T}}\psi_0 = \frac{\|\psi_0\|^2}{2}.$$

The theorem is proved. □

The following corollary to Theorem D-I.3 is of importance for constructing algorithms for pattern recognition.

Corollary. *If among the extreme vectors of a generalized portrait ψ_0 there are vectors of both classes, then the bound*

$$\rho(\psi_0) \leq \frac{1-k}{\sqrt{2W(\alpha, \beta)}} \tag{D-I.20}$$

is valid, where $\rho(\psi_0)$ is the distance between the projections of the sets $x_1, \ldots, x_a$ and $\bar{x}_1, \ldots, \bar{x}_b$ in the direction of a generalized portrait.

Moreover, equality in the bound (D-I.20) *is achieved at the point $\alpha = \alpha_0$, $\beta = \beta_0$.*

PROOF: In view of Theorem D-I.3,

$$\sqrt{2W(\alpha_0, \beta_0)} = \|\psi_0\|.$$

Furthermore, by virtue of the condition of the corollary there exist vectors in the set such that

$$\begin{aligned} x_i^{\mathrm{T}} \frac{\psi_0}{\|\psi_0\|} &= \frac{1}{\|\psi_0\|}, \\ \bar{x}_j^{\mathrm{T}} \frac{\psi_0}{\|\psi_0\|} &= \frac{k}{\|\psi_0\|}. \end{aligned} \tag{D-I.21}$$

Therefore the distance between projections of the vectors for which (D-I.21) is fulfilled equals

$$\rho(\psi_0) = \frac{1-k}{\|\psi_0\|} = \frac{1-k}{\sqrt{2W(\alpha_0, \beta_0)}}.$$

Taking into account that $W(\alpha, \beta) \leq W(\alpha_0, \beta_0)$, we obtain the inequality (D-I.20). □

This corollary is utilized for constructing a criterion for inseparability of vectors. Indeed, we shall assume that two finite sets of vectors cannot be subdivided by a hyperplane if the distance between the projections on the direction of the generalized portrait is less than ρ_0. This means that separability does not occur if one can find $\alpha^* > 0$, $\beta^* > 0$ such that

$$W(\alpha^*, \beta^*) > \frac{(1-k)^2}{2\rho_0^2} = W_0.$$

Thus when constructing a generalized portrait the problem is to find the maximum of a negative definite quadratic form $W(\alpha, \beta)$ in the positive quadrant $\alpha \geq 0, \beta \geq 0$ or to show that the maximum of the function $W(\alpha, \beta)$ exceeds W_0. The latter would indicate that it is impossible to construct a generalized portrait.

§3 Algorithms for Maximizing Quadratic Forms

One of the most efficient algorithms for the maximization of a negative definite quadratic form is the method of conjugate gradients. Using this method one can achieve the maximum in n steps, where n is the dimension of the form. In this section we shall consider algorithms for maximizing a negative definite quadratic form in the positive quadrant. These algorithms are based on a modification of the method of conjugate gradients. The theory of this method is described in various texts where a search for maximum values of functions is discussed (see, for example, [12, 80]).

Consider first the method of conjugate gradients applied to maximizing the quadratic form

$$F(y) = b^{\mathrm{T}}y - y^{\mathrm{T}}Ay,$$

where A is a positive definite matrix and b and y are vectors. According to the method, the search for the maximum starts from an arbitrary point $y_0 = y(0)$. The first step is taken in the direction of the gradient of $F(y)$ at the point $y(0)$. Denote the gradient of the function at $y(0)$ by $g(1)$, and the direction of the movement from $y(0)$ by $z(1)$. Thus

$$z(1) = g(1).$$

Steps are taken in the direction $z(1)$ until the maximum is attained. It is easy to verify that the maximum in the direction $z(1)$ is given by

$$y(1) = y(0) + \frac{z^{\mathrm{T}}(1)g(1)}{z^{\mathrm{T}}(1)Az(1)}\, z(1).$$

Starting with the second step the direction of the movement is determined by the vector

$$z(t+1) = g(t+1) + \frac{\|g(t+1)\|^2}{\|g(t)\|^2}\, z(t), \qquad \text{(D-I.22)}$$

where $g(t+1)$ and $g(t)$ are gradients of the function $F(y)$ at the points $y(t+1)$ and $y(t)$, and $z(t)$ is the direction of the movement at the point $y(t-1)$. The movement in the direction $z(t)$ is carried on until the conditional maximum is achieved. This maximum is attained at the point

$$y(t) = y(t-1) + h(t)z(t), \qquad \text{(D.I.23)}$$

where the quantity

$$h(z) = \frac{z^{\mathrm{T}}(t)g(t)}{z^{\mathrm{T}}(t)Az(t)}$$

determines the step of the movement.

The formulas (D-I.22) and (D-I.23) thus define the algorithm for searching for a maximum of a quadratic function $F(y)$.

To compute the maximum of a function in the positive quadrant we shall use the modified method of conjugate gradients. The modification of the method is designed to limit the region of search to the positive quadrant. We define the function

$$\hat{g}_i(y) = \begin{cases} \dfrac{\partial F(y)}{\partial y_i} & \text{if } y_i \neq 0 \text{ or } \dfrac{\partial F(y)}{\partial y_i} > 0, \\ 0 & \text{if } y_i = 0 \text{ and } \dfrac{\partial F(y)}{\partial y_i} \leq 0. \end{cases} \qquad \text{(D-I.24)}$$

The vector $\hat{g}(y) = (\hat{g}_1(y), \ldots, \hat{g}_n(y))^{\mathrm{T}}$ is the conditional gradient of the function $F(y)$ on the set $y \geq 0$.

We carry out our ascent towards the maximum utilizing formulas (D-I.22), (D-I.23) where $g(y)$ is replaced by $\hat{g}(y)$. The movement starts from an arbitrary point of the positive quadrant and continues until the moment at which the departure from the restriction at point y_0 takes place. Then the ascent starts again using the method of conjugate gradients, but this time from point y_0. The search for the maximum is terminated when the inequality

$$|\hat{g}_i(y)| \leq \varepsilon \qquad (i = 1, 2, \ldots, n)$$

is fulfilled. In order that the trajectory shall not depart from the positive quadrant, the size of the step $\hat{h}(t)$ is chosen to be the minimum of two quantities $h(t)$ and $h^*(t)$, where

$$h^*(t) = \min_i \frac{y_i(t)}{|z_i(t+1)|}.$$

When computing $h^*(t)$ the minimum is defined only for the coordinates y_i such that $z_i < 0$. If all $z_i \geq 0$, then the step equals $h(t)$.

An important special feature of this search method for the maximum of the function $F(y)$ in the positive quadrant is that it admits a sequential search procedure. Let the coordinates of the space E_n be

$$y_1, \ldots, y_k, y_{k+1}, \ldots, y_n.$$

One can first determine the conditional maximum of the function $F(y)$ under the restrictions

$$y_1 \geq 0, \ldots, y_k \geq 0, \qquad y_{k+1} = 0, \ldots, y_n = 0,$$

and then, using the obtained maximum point as the initial one, obtain the maximum of $F(y)$ in the region

$$y_1 \geq 0, \ldots, \qquad y_n \geq 0.$$

In our case, when searching for the maximum of the function

$$W(\alpha, \beta) = \sum_{i=1}^{a} \alpha_i - k \sum_{j=1}^{b} \beta_j - \tfrac{1}{2}\psi^{\mathrm{T}}\psi,$$

$$\psi = \sum_{i=1}^{a} \alpha_i x_i - \sum_{j=1}^{b} \beta_j \bar{x}_j,$$

in the positive quadrant, the conditional gradient is the vector with coordinates

$$\dot{\alpha}_i = \begin{cases} \dfrac{\partial W(\alpha, \beta)}{\partial \alpha_i} & \text{if } \alpha_i \geq 0 \text{ or } \dfrac{\partial W(\alpha, \beta)}{\partial \alpha_i} > 0, \\ 0 & \text{if } \alpha_i = 0 \text{ and } \dfrac{\partial W(\alpha, \beta)}{\partial \alpha_i} \leq 0, \end{cases}$$

$$i = 1, 2, \ldots, a,$$

$$\dot{\beta}_j = \begin{cases} \dfrac{\partial W(\alpha, \beta)}{\partial \beta_j} & \text{if } \beta_j \geq 0 \text{ or } \dfrac{\partial W(\alpha, \beta)}{\partial \beta_j} > 0, \\ 0 & \text{if } \beta_j = 0 \text{ and } \dfrac{\partial W(\alpha, \beta)}{\partial \beta_j} \leq 0, \end{cases}$$

$$j = 1, 2, \ldots, b.$$

Denote by $\bar{\alpha}$ and $\bar{\beta}$ the components of the vector $z(t)$ which determines the direction of the movement at the tth step. In view of (D-I.22) the relations

$$\begin{aligned} \bar{\alpha}(t+1) &= \dot{\alpha}(t+1) + \delta(t+1)\bar{\alpha}(t), \\ \bar{\beta}(t+1) &= \dot{\beta}(t+1) + \delta(t+1)\bar{\beta}(t) \end{aligned} \tag{D-I.25}$$

are fulfilled, where (cf. D-I.22)

$$\delta(t) = \frac{\sum_{i=1}^{a} \dot{\alpha}_i^2(t+1) + \sum_{j=1}^{b} \dot{\beta}_j^2(t+1)}{\sum_{i=1}^{a} \dot{\alpha}_i^2(t) + \sum_{j=1}^{b} \dot{\beta}_j^2(t)} \tag{D-I.26}$$

When computing a step by means of (D-I.23) it is necessary to compute the quantity $z^{\mathrm{T}}Az$. In our case

$$z^{\mathrm{T}}Az = \left(\sum_{i=1}^{a} \bar{\alpha}_i x_i - \sum_{j=1}^{b} \bar{\beta}_j \bar{x}_j \right)^2 = \bar{\psi}^{\mathrm{T}}\bar{\psi},$$

where we use the notation

$$\bar{\psi} = \sum_{i=1}^{a} \bar{\alpha}_i x_i - \sum_{j=1}^{b} \bar{\beta}_j \bar{x}_j.$$

Thus utilizing the congugate-gradient methods one can either determine a hyperplane which separates two sets of vectors, $x_1, \ldots, x_a$ and $\bar{x}_1, \ldots, x_b$ (i.e., determine the maximum of the function $W(\alpha, \beta)$ in the positive quadrant), or show that a separating hyperplane does not exist (i.e., establish that at the current step $W(\alpha, \beta) > (1 - k)^2/2\rho^2$, where ρ is a given parameter).

§4 Methods for Constructing an Optimal Separating Hyperplane

When devising algorithms for dependence estimation, one of the important steps is to construct an *optimal separating hyperplane*, that is, a hyperplane which subdivides the two sets of vectors $x_1, \ldots, x_a$ and $\bar{x}_1, \ldots, \bar{x}_b$, and is such that its distance to these vectors is maximal. Formally this means that an optimal separating hyperplane is defined by the following pair: a unit vector φ and a number c which satisfy

$$x_i^{\mathrm{T}}\varphi \geq c, \qquad i = 1, 2, \ldots, a,$$
$$x_j^{\mathrm{T}}\varphi < c, \qquad j = 1, 2, \ldots, b,$$

where

$$c = \frac{c_1(\varphi) + c_2(\varphi)}{2}, \qquad c_1(\varphi) = \min_i x_i^{\mathrm{T}}\varphi, \quad c_2(\varphi) = \max_j \bar{x}_j^{\mathrm{T}}\varphi,$$

and is such that the maximum of the expression

$$\Pi(\varphi) = c_1(\varphi) - c_2(\varphi)$$

is attained.

To construct an optimal separating hyperplane, consider all the differences

$$y_{ij} = x_i - \bar{x}_j; \qquad x_i \in X_a, \quad \bar{x}_j \in \bar{X}_b.$$

The vector φ_{opt} satisfies

$$\min_{i,j} y_{ij}^{\mathrm{T}}\varphi_{\mathrm{opt}} = \max_{\|\psi\|=1} \min_{i,j} y_{ij}^{\mathrm{T}} \frac{\psi}{\|\psi\|},$$

and hence is collinear with the minimal in the absolute value vector ψ for which the inequalities

$$y_{ij}^{\mathrm{T}}\psi \geq 1, \qquad i = 1, 2, \ldots, a, \qquad j = 1, 2, \ldots, b$$

are fulfilled. In other words, the vector φ is collinear with the generalized portrait ψ for the class $\{y_{ij}\}$ when the second class is empty.

One can determine the generalized portrait by maximizing the quadratic form

$$W(\alpha) = \sum_{i,j} \alpha_{ij} - \tfrac{1}{2}\psi^{\mathrm{T}}\psi,$$

$$\psi = \sum_{i,j} \alpha_{ij} y_{ij},$$

in the positive quadrant $\alpha_{ij} \geq 0$.

The number of vectors y_{ij} is usually quite large. Therefore the direct construction of a generalized portrait ψ is troublesome. We shall utilize the following iterative procedure:

(1) An arbitrary pair $y_1 = x_1 - \bar{x}_1$ is selected. A class Y_1 is formed consisting of a single vector $x_1 - \bar{x}_1$. A generalized portrait for this class is constructed (with an empty second class).

(2) Let the class Y_t of vectors $x_i - \bar{x}_i$ and its generalized portrait ψ_r be constructed at the tth step. In the training sequence there exists a vector $x_{i_{t+1}}$ such that

$$x_{i_{t+1}}^{\mathrm{T}}\psi_t = \min_{x_i \in X_a} x_i^{\mathrm{T}}\psi_t$$

and a vector $\bar{x}_{j_{t+1}}$ such that

$$\bar{x}_{j_{t+1}}^{\mathrm{T}}\psi_t = \max_{\bar{x}_i \in \bar{X}_b} \bar{x}_j^{\mathrm{T}}\psi_t.$$

The vector

$$y_{t+1} = x_{i_{t+1}} - \bar{x}_{j_{t+1}}$$

is then formed.

(3) If it turns out that

$$y_{t+1}^{\mathrm{T}}\psi_t < 1 - \varepsilon$$

(ε is the parameter of the algorithm), then the vector y_{t+1} is added to the class Y_t. A generalized portrait ψ_{t+1} of the class Y_{t+1} thus formed is then determined, and the process is continued further. If, however the inequality

$$y_{t+1}^{\mathrm{T}}\psi_t \geq 1 - \varepsilon$$

is fulfilled, then the process is completed and the hyperplane

$$x^{\mathrm{T}}\psi_t = \frac{\min_i x_i^{\mathrm{T}}\psi + \max_i \bar{x}_i^{\mathrm{T}}\psi}{2}$$

is selected as the optimal separating plane.

Simultaneously with the process of constructing the hyperplane the condition

$$W(\alpha) > \frac{2}{\rho^2}$$

is checked. If this condition is fulfilled even once, the construction of the hyperplane is terminated. In this case it is assumed that separation of the vectors of the training sequence is impossible.

In implementing this procedure it is expedient at each iteration when forming the class Y_t to omit vectors y_i which were included in the decomposition of the generalized portrait ψ with zero weight. A decrease in the number of vectors contained in Y_t allows us to shorten the time needed for constructing a generalized portrait ψ.

The algorithms for constructing the separating hyperplane considered above will be utilized for developing a battery of programs for pattern recognition. First, however, we shall discuss the question of representing information in recognition problems.

§5 An Algorithm for External Subdivision of Values of a Feature into Gradations

Two methods for representing information are used in the problem of pattern recognition: continuous and discrete. In the continuous method of representing information the coordinates of the vector x may take on arbitrary values. In the discrete method each coordinate of the vectors takes on a fixed number of values. The discrete method is suitable for coding qualitative features. For example, in problems of medical differential diagnostics, the features "paleness of the epidermis is not expressed", "moderately expressed", "strongly expressed" may be coded as 100, 010, 001.

However, in problems of pattern recognition it is customary to code in a discrete manner not only features which reflect qualitative indicators of an object, but also features which take on numerical values. Here the following method of representing information is utilized. The whole range of values of the parameter is subdivided into a number of *gradations*. The jth position of the code is coded 1 if the value of the parameter belongs to the jth gradation; otherwise the jth position is coded zero.

EXAMPLE. Let the value of the parameter x^i belong to the interval $[-5, 8]$, and this interval be subdivided into 5 gradations:

$$x^i < 0; \quad 0 \le x^i < 2; \quad 2 \le x^i < 4; \quad 4 \le x^i < 5; \quad x^i \ge 5.$$

The code 10000 denotes the values $x^i < 0$, the code 01000 denotes the values $0 \leq x^i < 2$, the code 00100 the values $2 \leq x^i < 4$, the code 00010 the values $4 \leq x^i < 5$, and the code 00001 the values $x^i \geq 5$.

This method of representing information is remarkable not only in that it allows us to write the information compactly (for the example presented above, instead of one memory cell in the computer, we need only five positions of a cell). Discretization of the coordinates of a vector is a nonlinear operation by means of which a vector x is transformed into a binary vector x' with a larger number of coordinates.

The utilization of a large number of gradations in coding a parameter is equivalent to the utilization of a more varied class of separating surfaces in the space E_n than that of linear surfaces. However, as was shown in Chapter 8, an excessively large capacity of the class of decision rules when the size of the training sequence is limited is inadmissible, and thus a problem of extremal subdivision into gradations of continuous features arises.

In this section we shall present an algorithm for extremal subdivision into gradations of values of a feature. The basic principle for implementing this algorithm is as follows: it is necessary to subdivide the values of the parameter into a finite number of gradations in such a manner that a measure of uncertainty (the entropy) in classification by means of this feature will be minimal (or close to minimal).

Thus let a feature (coordinate) x takes values in the interval $c \leq x \leq C$, and let the vector possessing this feature belong to one of K classes. Let there exist conditional probabilities

$$P(1|x), \ldots, P(K|x)$$

of belonging to each one of the classes. For each fixed value of the feature x a measure of the uncertainty (the entropy) of belonging to one of the K classes,

$$H(x) = -\sum_{i=1}^{K} P(i|x) \ln P(i|x)$$

is defined. The mean value of the entropy with respect to the measure $P(x)$ is computed as follows:

$$H = \int H(x)P(x)\,dx.$$

Now let the parameter x be subdivided into τ gradations, i.e., it takes one of the τ values $c(1), \ldots, c(\tau)$. Then the mean entropy can be written in the form

$$H(\tau) = -\sum_{j=1}^{\tau} \sum_{i=1}^{K} P(i|x_j) \ln P(i|x_j)P(x_j). \qquad \text{(D-I.27)}$$

We utilize Bayes's formula

$$P(i|x_j) = \frac{P(x_j|i)P(i)}{P(x_j)}, \tag{D-I.28}$$

Substituting (D-I.28) into (D-I.27), we have

$$H(\tau) = -\sum_{j=1}^{\tau}\sum_{i=1}^{K} P(x_j|i)P(i)\ln\frac{P(x_j|i)P(i)}{P(x_j)}. \tag{D.I.29}$$

In order to estimate the entropy (D-I.29) it is necessary to estimate the probabilities $P(x_j|i)$, $P(i)$, $P(x_j)$ from the training sequence. We shall utilize Bayes's estimators (see Section 6 of Chapter 3):

$$H(\tau) = -\sum_{j=1}^{\tau}\sum_{i=1}^{K}\frac{m_j(i)+1}{l(i)+\tau}\cdot\frac{l(i)+1}{l+K}\ln\left[\frac{m_j(i)+1}{l(i)+\tau}\cdot\frac{l(i)+1}{\sum_{i=1}^{K} m_j(i)+1}\cdot\frac{l+\tau}{l+K}\right],$$

where $l(i)$ is the number of elements in the ith class of the training sample, $m_j(i)$ is the number of vectors in the ith class for which $x = x_j$, and l is the sample size. Implementation of the formulated principle consists in choosing a subdivision of the interval $c \le x \le C$ into gradations such that the minimum of $H(\tau)$ will be achieved.

§6 An Algorithm for Constructing Separating Hyperplanes

In this section we shall consider two algorithms for constructing the separating hyperplane: the Special Algorithm and the General Algorithm.

The *Special Algorithm* is aimed at constructing a hyperplane which subdivides two finite sets of vectors or determining that an errorless linear subdivision of vectors is impossible.

This algorithm has two modifications. The first one determines a generalized portrait for a given parameter k, and the second determines the optimal separating hyperplane. The algorithm constructs the hyperplane by solving the dual problem of maximizing the quadratic form in the positive quadrant as it was described in Sections 3 and 4.

Modification 1: As stated above, this modification constructs the generalized portrait for a given k. Often, however the length of the training sequence is so large that in order to process the entire available training material one must solve a dual problem of exceedingly large dimensionality. Therefore the processing of the training sequence is carried out iteratively. The training sequence is subdivided into m groups with p elements in each group (the last group may be incomplete). Next a generalized portait for the vectors of the training sequence belonging to the first group is constructed. (The situation is favorable when the first group contains vectors belonging to the

first as well as to the second class.) This portrait is constructed by maximization of the corresponding quadratic form $W(\alpha, \beta)$ in the positive quadrant using the method of conjugate gradients (cf. Section 3). As a result of the maximization, either the generalized portrait is obtained or it is established that subdivision of the group of vectors which was set apart is impossible ($W(\alpha, \beta) > (1 - k)^2/2\rho^2$).

Let a generalized portrait ψ_1 be obtained from the first group. A working group of vectors is then formed, consisting of vectors appearing in the decomposition of a generalized portrait ψ_1 with nonzero weights and the vectors of the first two groups for which the inequalities

$$\begin{aligned} x_i^{\mathrm{T}}\psi_1 &< 1 - \delta, \\ \bar{x}_j^{\mathrm{T}}\psi_1 &> k + \delta \end{aligned} \tag{D.I.30}$$

are fulfilled, where δ is the parameter of the algorithm ($0 < \delta < (1 - k)/2$).

If in the first two groups no such vectors are available, then the working group of vectors is formed with the participation of the third group. If in the third group there are no vectors which satisfy the inequality (D-I.30), then the fourth group is considered and so on. If it turns out that in all m groups there are no vectors which satisfy (D-I.30), then ψ_1 is a generalized portrait for the whole training sequence.

Based on the constructed working group, the second iteration of the generalized portrait is carried out, and so on. The process is continued until either it happens that at some stage not a single vector is added to the formed group (which means that the generalized portrait has been constructed), or it has been established that an errorless separation of the vectors of the training sequence using the hyperplane is impossible.

Modification 2: This modification of the algorithm is designed for constructing an optimal separating hyperplane. This also is done iteratively.

For the first iteration a working group of vectors is formed, consisting of l_1 vectors $x_1, \ldots, x_{l_1}$ of the training sequence belonging to the first class and l_1 vectors $\bar{x}_1, \ldots, \bar{x}_{l_1}$ of the training sequence belonging to the second class. Using these vectors, l vectors $y_i = x_i - \bar{x}_i$, $i = 1, \ldots, l$, are formed, for which a generalized portrait (of one class with an empty second class) is sought. This generalized portrait is determined by means of maximization of the quadratic form $W(\alpha)$ in the positive quadrant (cf. Section 4).

Let the generalized portrait ψ_1 be determined as a result of the first iteration. To obtain the second iteration we form a working group of differences Y_2. To do this we omit from the working group of differences Y_1 those pairs which appeared in the expansion of ψ_1 with zero weights, and find among the vectors of the training sequence vectors x_i and $\bar{x}_j$ for which the extremal values

$$\begin{aligned} x_*^{\mathrm{T}}\psi_1 &= \min_i x_i^{\mathrm{T}}\psi_1, \\ \bar{x}_*^{\mathrm{T}}\psi_1 &= \max_j \bar{x}_j^{\mathrm{T}}\psi_1 \end{aligned}$$

are attained.

Suppose it turns out that the inequalities

$$\begin{aligned} x_*^T\psi_1 &\geq \min_x x^T\psi_1 - \delta_1, \\ \bar{x}_*^T\psi_1 &\leq \max_x \bar{x}^T\psi_1 + \delta_2 \end{aligned} \tag{D-I.31}$$

are fulfilled, where on the right-hand sides the minimum and the maximum are computed only over the vectors of the training sequence which appear in the working group, and δ_1 and δ_2 are parameters of the algorithm (usually $\delta_1 = 0.1\ (\min_i x_i^T\psi)$, $\delta_2 = 0.1\ (\max_j \bar{x}_j^T\psi)$). Then the pair consisting of the vector ψ and the number

$$\left(\frac{\min_i x_i^T\psi + \max_j x_j^T\psi}{2}\right) = c$$

defines the optimal separating hyperplane. If, on the other hand, at least one of the inequalities (D-I.31) is not fulfilled, then the pair x^*, $\bar{x}^*$ is added to the working group of vectors and a new iteration of the optimal separating hyperplane is constructed. The process continues until either both inequalities are violated at one time or it turns out that the separation is impossible ($W(\alpha) > 2/\rho^2$, where ρ is a given number).

Thus using the Special Algorithm one is able either to construct a separating hyperplane or to establish that an errorless separation of the vectors of the training sequence is impossible.

In accordance with the bound (D-I.2), if it is possible, in a space of dimension n, to construct a hyperplane which errorlessly separates l vectors of the training sequence, then one can assert with probability $1 - \eta$ that the probability of erroneous classification by means of the constructed hyerplane will be less than

$$P < \frac{n\left(\ln\frac{l}{n} + 1\right) - \ln\eta}{l}.$$

The *General Algorithm* is designed for constructing a hyperplane which separates two sets of vectors with a minimal number of errors.

The problem of constructing a separating hyperplane which minimizes the number of incorrectly classified vectors can be solved in principle by solving the problem of constructing a separating hyperplane; however, the precise solution of the latter problem requires a large number of enumerations. We shall therefore apply a heuristic method which permits us to reduce this number.

The General Algorithm utilizes the following heuristic device: from the set of vectors of the training sequence a single element is excluded which "hinders separation to the largest extent"; next—provided the subdivision is impossible—from the remaining set yet another element is excluded and so on. A special characteristic of this algorithm consists in the definition of

the element which "hinders separation to the largest extent". In constructing a generalized portrait, we choose for this element the vector x_i (or $\bar{x}_j$) that at the stopping time yields the largest contribution to the value of

$$W(\alpha, \beta) = \sum_{i=1}^{a} \alpha_i(1 - \tfrac{1}{2}x_i^T\psi) + \sum_{j=1}^{b} \beta_j(-k + \tfrac{1}{2}\bar{x}_i^T\psi).$$

In other words, the vector x_i ($\bar{x}_j$) for which the maximum of the quantity

$$\alpha_i(1 - \tfrac{1}{2}x_j^T\psi) \qquad (\beta_j(-k + \tfrac{1}{2}x_j^T\psi))$$

is attained is selected as the "most hindering vector". (In Modification 2 $k = 1$.)

The General Algorithm excludes from the training sequence the most hindering vector, separates the remaining set of vectors, and, if separation is still impossible, again excludes a vector, separates the remaining set, and so on.

Finally, after m vectors have been excluded, The General Algorithm separates the remaining set of vectors, constructing the separating hyperplane $x^T\psi = c$. In accordance with the bound D-I.2, with probability $1 - \eta$ the error of classification using the constructed hyperplane is bounded by

$$P < \frac{n\left(\ln \frac{l}{n} + 1\right) - \ln \eta}{2l}\left(1 + \sqrt{1 + \frac{4m}{n\left(\ln \frac{l}{n} + 1\right) - \ln \eta}}\right) + \frac{m}{l}.$$

The General Algorithm for constructing the separating hyperplane is basic for all the pattern recognition algorithms described in this book:

(1) Algorithm for constructing the hyperplane in the optimal feature space (Chapter 8, Section 5).
(2) Algorithm for the estimation of function values at the given points in the class of linear separating hyperplanes (Chapter 10, Section 5).
(3) Algorithm for the estimation of function values in the class of piecewise linear separating hyperplanes (Chapter 10, Section 9).
(4) Algorithm for the estimation of function values in the class of locally linear separating hyperplanes (Chapter 10, Section 11).

Special realizations of each of these algorithms are determined by the schemes chosen for solving the corresponding problems of discrete optimization.

Addendum II

Algorithms for Estimating Nonindicator Functions

§1 Remarks Concerning Algorithms

In this Addendum algorithms for estimating nonindicator functions are considered. As above, two problems of estimation will be distinguished: estimation of functional dependence and estimation of values of a function at given points.

The two bounds obtained in Chapters 8 and 10 serve as the basis of the algorithms considered herein. The first bound,

$$I(\alpha) < \left[\frac{I_{\text{emp}}(\alpha)}{1 - 2\tau a(p) \sqrt{\dfrac{h\left(\ln \dfrac{2l}{h} + 1\right) - \ln \dfrac{\eta}{12}}{l}}} \right]_{\infty}, \qquad \text{(D-II.1)}$$

connects the value of the expected risk $I(\alpha)$ with the value of the empirical risk $I_{\text{emp}}(\alpha)$; the second bound,

$$I_{\Sigma}(\alpha) < \left[\frac{1 + \tau a(p) \dfrac{l}{l+k} \varkappa_*}{1 - \tau a(p) \dfrac{k}{l+k} \varkappa_*} \right]_{\infty} I_{\text{emp}}(\alpha), \qquad \text{(D-II.2)}$$

where $\varkappa_*$ is the smallest solution of the inequality

$$h\left(\ln \frac{l+k}{h} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leq \ln \frac{\eta}{1.5}, \qquad \text{(D-II.3)}$$

connects the value $I_{\Sigma}(\alpha)$ of the overall risk at points of the working sample with the value of the empirical risk $I_{\text{emp}}(\alpha)$.

Unlike the analogous estimates utilized in the estimation of indicator dependences, the bounds (D-II.1) and (D-II.2) contain a free parameter τ. According to the theory this parameter determines a statistical characteristic of the problem (an allowable value of a possible large deviation), and its value should be known *a priori*.

Below we shall utilize bounds suitable for "real world" situations where specific values for constants are given. We shall use the bound

$$I(\alpha) < \left[\frac{I_{\text{emp}}(\alpha)}{1 - \sqrt{\frac{h\left(\ln \frac{l}{h} + 1\right) - \ln \eta}{l}}} \right]_{\infty} \tag{D-II.4}$$

for estimating the functional dependence, and the bound

$$I_{\Sigma}(\alpha) < \left[\frac{1 + \frac{1}{l+k} \varkappa_*}{1 - \frac{k}{l+k} \varkappa_*} \right]_{\infty} I_{\text{emp}}(\alpha), \tag{D-II.5}$$

where $\varkappa_*$ is the smallest solution of the inequality

$$h\left(\ln \frac{l+k}{h} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leq \ln \eta \tag{D-II.6}$$

for estimating values of a function at given points.

§2 An Algorithm for Regression Estimation in a Class of Polynomials

Consider algorithms for the estimation of one-dimensional (univariate) functional dependence based on emprical data

$$x_1, y_1; \ldots; x_l, y_l$$

in a class of linear (in parameters) functions

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x).$$

We shall assume that the functions

$$\varphi_1(x), \ldots, \varphi_n(x) \tag{D-II.7}$$

are ordered *a priori*, i.e., the structure

$$S_1 \subset S_2 \subset \cdots \subset S_n \tag{D-II.8}$$

is defined, where the element S_p is the set of functions

$$F(x, \alpha) = \sum_{i=1}^{p} \alpha_i \varphi_i(x).$$

In this case the problem is reduced to the determination of an element S_p of the structure (D-II.8) and of a function $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk in S_p, so that the minimum of the functional

$$R(p) = \left[\frac{I_{\text{emp}}(\alpha_{\text{emp}})}{1 - \sqrt{\dfrac{p\left(\ln \dfrac{l}{p} + 1\right) - \ln \eta}{l}}} \right]_{\infty} \qquad (p < l)$$

is attained. The minimum of the empirical risk

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{j=1}^{p} \alpha_j \varphi_j(x_i) \right)^2 \qquad \text{(D-II.9)}$$

in S_p is computed by means of standard methods of linear algebra: the vector of parameters $\alpha_{\text{emp}} = (\alpha_1^{\text{emp}}, \ldots, \alpha_p^{\text{emp}})^{\mathrm{T}}$ is equal to

$$\alpha_{\text{emp}} = (\Phi_p^{\mathrm{T}} \Phi_p)^{-1} \Phi_p^{\mathrm{T}} Y, \qquad \text{(D-II.10)}$$

where Y is the vector of values $y_1, \ldots, y_l$ and Φ_p is the matrix

$$\Phi_p = \begin{Vmatrix} \varphi_1(x_1) & \cdots & \varphi_p(x_1) \\ \vdots & & \vdots \\ \varphi_1(x_l) & \cdots & \varphi_p(x_l) \end{Vmatrix}. \qquad \text{(D-II.11)}$$

The problem of inverting matrices of the type $(\Phi^{\mathrm{T}}\Phi)$ has been studied extensively (see e.g., [63], [59]). Any algorithm for inversion recommended in those references may be used.

Thus the only problem which arises in implementing the scheme under consideration is to decide which one of the specific systems of functions (D-II.7) should be utilized.

We estimate the function in the class of polynomials, i.e., we assume that $\varphi_p(x)$ is a polynomial of degree $p - 1$:

$$\varphi_p(x) = \sum_{s=1}^{p} \beta_s x^{s-1}.$$

In principle it is irrelevant how the polynomials $\varphi_p(x)$ are defined (as long as the coefficients at the highest degrees do not vanish). Therefore it is often assumed that $\varphi_p(x) = x^{p-1}$.

Thus *Algorithm* D-II.1 for estimating one-dimensional functional dependences in a class of polynomials has been determined.

It was shown in Chapter 9 that a close approximation of the desired function in a class of polynomials can be guaranteed only in the integral sense, while in a class of piecewise polynomial dependences one can achieve not

only integral approximation but uniform approximation on the whole interval of definition of the function. It turns out (this will be shown in the next section) that the problem of constructing approximations for functions in a class of piecewise polynomial dependences is slightly more involved in its computational aspect than that of approximating in a class of polynomials.

§3 Canonical Splines

Let the interval $[a, b]$ on which the estimation of dependences is carried out be subdivided into $N + 1$ parts

$$[a_0, a_1), [a_1, a_2), \ldots, [a_N, b].$$

Consider the following class of functions: on each of the $N + 1$ subintervals, each function coincides with a polynomial of degree m (different polynomials on different subintervals) and is continuous on the whole interval together with its first $m - 1$ derivatives. Such a class of functions is called a *class of splines* of degree m conjugated at N points $a_1, \ldots, a_N$.

Below we shall assume that $m = 3$ and the points $a_1, \ldots, a_N$ which define the subintervals are obtained from the subdivision of the interval $[a, b]$ into $N + 1$ equal parts. The class of such splines will be denoted by $V_N^3(x, \alpha)$.

The problem is to find a function $f_N^3(x, \alpha_{\text{emp}})$ belonging to $V_N^3(x, \alpha)$ which minimizes the empirical risk

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - V_N^3(x_i, \alpha))^2. \tag{D-II.12}$$

It is convenient to construct splines by introducing a system of canonical splines. For cubic splines with N conjugations on the grid $(a, a_1, \ldots, a_N, a_{N+1} = b)$, $N + 4$ canonical splines are introduced:

$$\mu_1(x), \mu_2(x), \mu_3(x), \ldots, \mu_{N+4}(x). \tag{D-II.13}$$

The canonical splines (D-II.13) are uniquely defined by the conditions

$$\mu_1(a_i) = 0, \quad \mu_1'(\alpha_0) = 1, \quad \mu_1'(a_{N+1}) = 0 \qquad (i = 1, 2, \ldots, N + 1),$$

$$\mu_2(a_i) = 0, \quad \mu_2'(a_0) = 0, \quad \mu_2'(a_{N+1}) = 1 \qquad (i = 1, 2, \ldots, N + 1).$$

$$\mu_r(a_k) = \delta_{k, r-3}, \quad \mu_r'(\alpha_0) = \mu_r'(a_{N+1}) = 0$$

$$(r = 3, \ldots, N + 4; \quad k = 0, 1, \ldots, N + 1),$$

$$a_0 = a, \qquad a_{N+1} = b.$$

Here δ_{ij} is the Kronecker symbol:

$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$

Since any spline $V_N^3(x, \alpha)$ is completely determined by the $N + 2$ values at the nodes a_i $(i = 0, 1, \ldots, N + 1)$ and the values of the first derivative at the endpoints of the interval, the equality

$$V_N^3(x, \alpha) = \sum_{j=0}^{N+1} V_N^3(a_j, \alpha)\mu_{j+3}(x) + [V_N^3(x, \alpha)]'_a \mu_1(x) + [V_N^3(x, \alpha)]'_b \mu_2(x)$$

is valid. We shall utilize this representation below when estimating regression in the class of splines $V_N^3(x, \alpha)$. Moreover, we shall obtain specific expressions for the system of canonical splines $\mu_1(x), \mu_2(x), \mu_3(x), \ldots, \mu_{N+4}(x)$, and using this system we shall represent the class of splines of degree 3 with N conjugations in the parametric form

$$V_r^3(x, \alpha) = \sum_{i=1}^{N+4} \alpha_i \mu_i(x).$$

We shall thus reduce the problem of determining a spline minimizing the empirical risk (D-II.11) to the determination of parameters α minimizing the functional

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{j=4}^{N+4} \alpha_j \mu_j(x_i) \right)^2,$$

i.e., we shall reduce the solution of the problem to the same linear system (D-II.10) which determined estimation of regression in the class of polynomials. We thus construct a system of cubic canonical splines on a uniform grid with the step Δ: $a_{i+1} - a_i = \Delta$.

Let m_{i+1}, m_i be the values of the second derivative of the spline $V_N^3(x, \alpha)$ at the nodes a_{i+1} and a_i. Since the second derivative of a third-degree polynomial is a linear function, the equality

$$[V_N^3(x, \alpha)]'' = m_{i+1} \frac{x - a_i}{\Delta} + m_i \frac{a_{i+1} - x}{\Delta},$$

where

$$m_{i+1} = [V_N^3(x, \alpha)]''_{a_{i+1}}, \qquad m_i = [V_N^3(x, \alpha)]''_{a_i}$$

is valid for all $x \in [a_i, a_{i+1}]$.

Integrating this function twice and taking into account the continuity of a spline at the endpoints of the interval $[a_i, a_{i+1}]$, we obtain the following representation for a cubic spline on the interval $[a_i, a_{i+1}]$:

$$V_N^3(x, \alpha) = \frac{1}{6\Delta} [m_i(a_{i+1} - x)^3 + m_{i+1}(x - a_i)^3 + (6V_N^3(a_i, \alpha) - \Delta^2 m_i)(a_{i+1} - x) + (6V_N^3(a_{i+1}, \alpha) - \Delta^2 m_{i+1})(x - a_i)].$$

The function obtained is continuous on the whole interval $[a, b]$, but its first derivative may have discontinuities at the nodes of conjugations. To

avoid these discontinuities we shall choose the values of m from the condition of continuity of the derivative of the spline on the whole interval $[a, b]$. Equating one-sided derivatives of the spline at points a_{i+1}, we obtain the equations

$$[V_N^3(x, \alpha)]'_{a_{i+1}-0} = \frac{\Delta}{3} m_i + \frac{\Delta}{3} m_{i+1} + \frac{V_N^3(a_{i+1}, \alpha) - V_N^3(a_i, \alpha)}{\Delta}$$

$$= -\frac{\Delta}{3} m_{i+1} - \frac{\Delta}{3} m_{i+2} + \frac{V_N^3(a_{i+2}, \alpha) - V_N^3(a_{i+1}, \alpha)}{\Delta}$$

$$= [V_N^3(x, \alpha)]'_{a_{i+1}+0}.$$

Thus we have obtained N linear equations for the determination of $N + 2$ values m_i. The boundary conditions

$$[V_N^3(x, a)]'_a = V'_a, \qquad [V_N^3(x, a)]'_b = V'_b$$

supply an additional two equations; hence we arrive at

$$2m_1 + m_2 = \frac{6}{\Delta}\left(\frac{V_N^3(a_1, \alpha) - V_N^3(a_0, \alpha)}{\Delta} - V'_a\right),$$

$$m_{N+1} + 2m_{N+2} = \frac{6}{\Delta}\left(V'_b - \frac{V_N^3(a_{N+1}, \alpha) - V_N^3(a_N, \alpha)}{\Delta}\right).$$

In the matrix notation the system becomes

$$\mathscr{C}\,\mathscr{M}^* = \mathscr{D},$$

where

$$\mathscr{C} = \begin{pmatrix} 2 & 1 & 0 & \cdots & & & & & \\ \frac{1}{2} & 2 & \frac{1}{2} & 0 & \cdots & & & & \\ 0 & \frac{1}{2} & 2 & \frac{1}{2} & 0 & \cdots & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & \cdots & & & 0 & \frac{1}{2} & 2 & 0 & \\ & \cdots & & & & 0 & \frac{1}{2} & 2 & \frac{1}{2} \\ & \cdots & & & & & 0 & 1 & 2 \end{pmatrix}$$

$$\mathscr{M}^* = \begin{bmatrix} m_1 \\ \vdots \\ m_{N+2} \end{bmatrix}, \qquad \mathscr{D} = \begin{bmatrix} d_1 \\ \vdots \\ d_{N+2} \end{bmatrix},$$

$$d_1 = \frac{6}{\Delta}\left(\frac{V_N^3(a_1, \alpha) - V_N^3(a_0, \alpha)}{\Delta} - V'_a\right),$$

$$d_i = \frac{3}{\Delta^2}\left(V_N^3(a_{i+2}, \alpha) + V_N^3(a_i, \alpha) - 2V_N^3(a_{i+1}, \alpha)\right), \qquad i = 2, \ldots, N,$$

$$d_{N+2} = \frac{6}{\Delta}\left(V'_b - \frac{V_N^3(a_{N+1}, \alpha) - V_N^3(a_N \alpha)}{\Delta}\right).$$

To construct canonical splines $\mu_1(x), \ldots, \mu_{N+4}(x)$ it is convenient to represent the vector $\mathscr{D} = (d_1, \ldots, d_{N+2})$ as the product of the vector of defining values

$$V_i = ([V_N^3(a_0, \alpha)]', V_N^3(a_0, \alpha), \ldots, V_N^3(a_{N+1}, \alpha), [V_N^3(a_{N+1}, \alpha)]')^{\mathrm{T}}$$

and the matrix $\mathscr{B}$ given by

$$\mathscr{B} = \begin{bmatrix} -\frac{6}{\Delta} & -\frac{6}{\Delta^2} & \frac{6}{\Delta^2} & 0 & \cdots & & \\ 0 & \frac{3}{\Delta^2} & -\frac{6}{\Delta^2} & \frac{3}{\Delta^2} & 0 & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & \cdots & 0 & \frac{3}{\Delta^2} & -\frac{6}{\Delta^2} & \frac{3}{\Delta^2} & 0 \\ & & & \cdots & 0 & \frac{6}{\Delta^2} & -\frac{6}{\Delta^2} & \frac{6}{\Delta} \end{bmatrix}.$$

The vectors V_i^*, which have $N + 3$ coordinates equal to zero and one coordinate equal to 1, serve as the defining values for canonical splines. The location of the 1 in the vector is determined by the ordinal number of the canonical spline. Under a suitable ordering of canonical splines the matrix of defining values becomes the unit matrix. Below we present such an ordering:

$$\begin{array}{c} \begin{matrix} \mu_1 & \mu_3 & & \mu_{N+4} & \mu_2 \end{matrix} \\ \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \end{array}.$$

The matrix $\mathscr{M}$ of values of second derivatives of $N + 4$ canonical splines,

$$\mathscr{M} = \begin{bmatrix} m_{1,1} & \cdots & m_{1,N+4} \\ \cdots & \cdots & \cdots \\ m_{N+2,1} & \cdots & m_{N+2,N+4} \end{bmatrix},$$

is determined as the solution of the matrix equation

$$\mathscr{C}\mathscr{M} = \mathscr{B}.$$

Knowing the matrix $\mathcal{M}$, it is easy to compute the values of canonical splines. These are computed using the formulas: for $x \in [a_i, a_{i+1}]$

$$\mu_1(x) = m_{i1}\frac{(a_{i+1}-x)^3}{6\Delta} + m_{i+1,1}\frac{(x-a_i)^3}{6\Delta}$$

$$- m_{i+1,1}\frac{x-a_i}{6}\Delta - m_{i,1}\frac{a_{i+1}-x}{6}\Delta,$$

$$\mu_2(x) = m_{i,N+4}\frac{(a_{i+1}-x)^3}{6\Delta} + m_{i+1,N+4}\frac{(x-a_i)^3}{6\Delta}$$

$$- m_{i+1,N+4}\frac{x-a_i}{6}\Delta - m_{i,N+4}\frac{a_{i+1}-x}{6}\Delta,$$

$$\mu_j(x) = m_{i,j-2}\frac{(a_{i+1}-x)^3}{6\Delta} + m_{i+1,j-2}\frac{(x-a_i)^3}{6\Delta}$$

$$+ \left(\delta_{i,j-3} - m_{ij-2}\frac{\Delta^2}{6}\right)\frac{a_{i+1}-x}{\Delta}$$

$$+ \left(\delta_{i+1,j-3} - m_{i+1,j-2}\frac{\Delta^2}{6}\right)\frac{x-a_i}{\Delta}.$$

The matrix $\mathcal{M}$ may be computed explicitly:

$$\mathcal{M} = \mathcal{C}^{-1}\mathcal{B}.$$

(For this purpose it is sufficient to obtain the matrix $\mathcal{C}^{-1}$; see below.)

Denote by D_n the determinant of order n:

$$D_n = \det\begin{bmatrix} 2 & 1 & 0 & & & 0 \\ \frac{1}{2} & 2 & \frac{1}{2} & 0 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & 0 & \frac{1}{2} & 2 & \frac{1}{2} \\ 0 & & & 0 & \frac{1}{2} & 2 \end{bmatrix}.$$

Expanding this determinant with respect to the cofactors of the elements of the last column, we arrive at a recursive formula for computing the determinant D_n:

$$D_n = 2D_{n-1} - \tfrac{1}{2}D_{n-2}, \qquad D_1 = 2, \quad D_0 = 1.$$

We shall now evaluate the elements c_{ij}^{-1} of the matrix $\mathscr{C}^{-1}$ utilizing the cofactors of matrix $\mathscr{C}$ expressed by means of the determinant D_n. We obtain

$$\begin{aligned}
&\text{I:} \quad c_{ij}^{-1} = \frac{(-1)^{i+j}}{2^{j-i}} \frac{D_{i-1} D_{N+2-j}}{D_{N+2}} \qquad (1 < i \le j \le N+2),\\
&\text{II:} \quad c_{ij}^{-1} = \frac{(-1)^{i+j}}{2^{i-j}} \frac{D_{j-1} D_{N+2-i}}{D_{N+2}} \qquad (1 \le j \le i < N+2),\\
&\text{III:} \quad c_{1,j}^{-1} = \frac{(-1)^{1+j}}{2^{j-2}} \frac{D_{N+2-j}}{D_{N+2}} \qquad (1 < j \le N+2),\\
&\text{IV:} \quad c_{N+2,j}^{-1} = \frac{(-1)^{N+1}}{2^{N+1-i}} \frac{D_{j-1}}{D_{N+2}} \qquad (1 \le j < N+2).
\end{aligned} \tag{D-II.14}$$

The scheme for application of formulas I–IV can vividly be represented graphically (Figure 22).

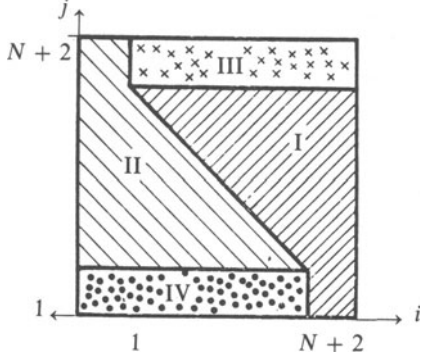

Figure 22

Thus in order to obtain a system of cubic canonical splines with N conjugates one is required:

(1) to compute the values $\mathscr{D}_n$;
(2) using (D-II.14), to obtain the matrix $\mathscr{C}^{-1}$ (of dimension $(N+2) \times (N+2)$);
(3) to compute the matrix $\mathscr{M}$ (of dimension $(N+2) \times (N+4)$), by multiplying $\mathscr{C}^{-1}$ by $\mathscr{B}$ (the matrix $\mathscr{B}$ is of dimension $(N+2) \times (N+4)$);
(4) using the formulas (D-II.14), to obtain the canonical splines.

In order to retain uniform notation we shall use the same symbols to denote systems of canonical splines as those used to denote systems of polynomials, i.e., we set

$$\mu_1(x) = \varphi_1(x), \ldots, \qquad \mu_{N+4}(x) = \varphi_{N+4}(x).$$

Using this notation the problem of determining a cubic spline minimizing the empirical risk (D-II.12) will be written in the form (D-II.9). Its solution—the determination of the vector α of coefficients of expansion of the required functions in terms of a system of canonical splines—is given by (D-II.10). Thus after the canonical system of splines has been constructed, the evaluation of the spline which minimizes the empirical risk is carried out using exactly the same procedure which was used to determine the coefficients of a linear (in parameters α) regression.

§4 Algorithms for Estimating Functions in a Class of Splines

We now present *Algorithm* D-II.2 for structural minimization of the risk in a class of splines. For this purpose we shall define the following structure on the set of piecewise polynomial dependencies. The class S_1 will contain all the constants; the class S_2, of all the polynomials of the first degree; the class S_3, polynomials of the second degree; and the class S_4, polynomials of the third degree (we shall call them cubic splines with zero conjugates).

Starting with the fifth class, piecewise polynomial functions are considered. The fifth class S_5 contains splines with one conjugate, S_6 those with two, and so on.

The capacity of the set of functions formed by splines with r conjugates is equal to $h = r + 4$.

Thus the problem is to choose an element of the structure S_{r+4} for which the minimum—with respect to α and r—of the functional

$$R(\alpha, r) = \left[\frac{\frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{p=1}^{r+4} \alpha_p \varphi_p(x_i) \right)^2}{1 - \sqrt{\frac{(r+4)\left(\ln \frac{l}{r+4} + 1 \right) - \ln \eta}{l}}} \right]_{\infty} \qquad \text{(D-II.15)}$$

is attained.

A special feature of the problem of estimating regression in the class of piecewise polynomial dependences is that each time we advance to a class of splines with a larger number of conjugates, a new canonical system (its own) is used. (Recall that in a similar situation of estimating regression in a class of polynomials we simply add a new function to the system.) Strictly speaking, this causes the element S_{p+1} of the structure not to contain S_p. However, this fact is not of basic importance in this case.

When estimating nonindicator functional dependence, it is expedient to carry out a selection of the training sequence, i.e., to exclude a number $t = 0$,

1, 2, ... of vectors such that the functional

$$\hat{R}(\alpha, r) = \left[\frac{\frac{1}{l-t} \sum_{i}^{(t)} \left(y_i - \sum_{p=1}^{r+4} \alpha_p \varphi_p(x_i) \right)^2}{1 - \sqrt{\frac{h\left(\ln \frac{l-t}{h} + 1 \right) - \ln \eta + \ln C_l^t}{l-t}}} \right]_{\infty} \tag{D-II.16}$$

will attain its "deepest" minimum. The function for which (D-II.16) achieves its minimum is chosen for the solution of the problem of minimizing the expected risk ($\sum_{i}^{(t)}$ denotes that only $l - t$ terms are summed).

Determination of the exact minimum for the functional (D-II.16) requires a large number of enumerations. Therefore it seems reasonable to use here the method of successive decrease of a functional. First we determine a vector whose exclusion from the training sequence will minimize the functional for $t = 1$. If this quantity turns out to be smaller than the minimal value of the functional (D-II.16) for $t = 0$ (for the whole training sample), then the corresponding vector is deleted and an attempt is made to analogously exclude yet another vector, i.e., to minimize (D-II.16) for $t = 2$, and so on. If no exclusion of a single vector results in a decrease of the functional, then the exclusion process is terminated.

§5 Algorithms for Solving Ill-posed Problems of Interpreting Measurements

In this section we shall consider algorithms for solving ill-posed problems of interpreting results of indirect experiments in the case when the operator equation

$$Af(t) = F(x) \tag{D-II.17}$$

is a Fredholm integral equation of the type I:

$$\int_a^b K(t, x) f(t)\, dt = F(x). \tag{D-II.18}$$

Let measurements on a function $F(x)$ at l points x_i be given:

$$x_1, y_1; \ldots; x_l, y_l.$$

According to the theory, the function $f(t)$ which yields the minimum of the functional

$$I = \int \left(y - \int_a^b K(t, x) f(t)\, dt \right)^2 P(y|x)\, dy\, dx \tag{D-II.19}$$

is a solution of Equation (D-II.18).

We shall minimize the expected risk (D-II.19) using the method of structural minimization in the class of cubic splines. To do this we shall determine a function $V_r^3(t, \alpha_{\text{emp}})$ which minimizes the following functional in the class of cubic splines with r conjugations:

$$R(\alpha, r) = \left[\frac{\frac{1}{l} \sum_{i=1}^{l} \left(y_i - \int_a^b K(t, x_i) V_r^3(t, \alpha)\, dt \right)^2}{1 - \sqrt{\frac{(r+4)\left(\ln \frac{l}{r+4} + 1 \right) - \ln \eta}{l}}} \right]_+ \tag{D-II.20}$$

We construct *Algorithm* D-II.3 for minimization of the functional (D-II.20) as a modification of Algorithm D-II.2. For this purpose we introduce the notation

$$\int_a^b K(t, x_i)\, \mu_p(t)\, dt = \varphi_p(x). \tag{D-II.21}$$

Since, in view of Section 3,

$$V_r^3(t, \alpha) = \sum_{p=1}^{r+4} \alpha_p \mu_p(t),$$

using the notation (D-II.21), the minimization of the functional (D-II.20) is reduced to that of the functional (D-II.15). The minimum of the functional with respect to α and r may be obtained using the scheme of Algorithm D-II.2. Let the minimum be attained at r^* and α_{emp}. Then the function

$$f(t) = \sum_{p=1}^{r^*+4} \alpha_p^{\text{emp}} \mu_p(t)$$

is declared to be a solution of the integral equation. When interpreting results of indirect experiments it is desirable to carry out a selection of observations. This selection also is carried out following Algorithm D-II.2, i.e., it is reduced to minimizing the functional (D-II.16) and then choosing as the solution the preimage of the function which yields the minimum for this functional.

§6 Algorithms for Estimating Multidimensional Regression in a Class of Linear Functions

We now present *Algorithm* D-II.4 for estimating multidimensional linear regression. It is required to estimate regression in the class of functions

$$F(x, \beta) = \sum_{i=1}^{n} \beta_i \varphi_i(x) = \beta^{\mathrm{T}} \varphi(x)$$

$$(\varphi(x) = (\varphi_1(x), \ldots, \varphi_n(x))^{\mathrm{T}}). \tag{D-II.22}$$

Let

$$\zeta_1, \ldots, \zeta_n \tag{D-II.23}$$

be a system of eigenvectors of the matrix $(\Phi^T\Phi)$ (the matrix (D-II.11)) ordered according to the decreasing value of the eigenvalues. We represent (D-II.22) in the form

$$F(x, \alpha) = \sum_{p=1}^{n} \alpha_p \zeta_p^T \varphi(x) = \sum_{p=1}^{n} \alpha_p \chi_p(x), \tag{D-II.24}$$

where

$$\chi_p(x) = \zeta_p^T \varphi(x).$$

Define the structure

$$S_1 \subset \cdots \subset S_n \tag{D-II.25}$$

on the class of functions $F(x, \alpha)$, where S_p contains only those functions which can be expanded in terms of the first p members of the series.† Then the best element of the structure will be the one for which the minimum of the functional

$$R(\alpha, p) = \left[\frac{\frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{s=1}^{p} \alpha_s \chi_s(x_i) \right)^2}{1 - \sqrt{\frac{p\left(\ln \frac{l}{p} + 1 \right) - \ln \eta}{l}}} \right]_{\infty}$$

is attained. The implementation of this algorithm is the same as that of Algorithm D-II.2.

When constructing linear regression it often turns out to be desirable to carry out a selection of the training sequence. It is necessary to exclude t elements ($t = 0, 1, 2, \ldots, s$) such that the functional

$$R(\alpha, p) = \left[\frac{I_{\text{emp}}^t(\alpha)}{1 - \sqrt{\frac{p\left(\ln \frac{l-t}{p} + 1 \right) + \ln C_l^t - \ln \eta}{l - t}}} \right]_{\infty} \tag{D-II.26}$$

will be minimized with respect to α and p. Here $I_{\text{emp}}^t(\alpha)$ is an empirical-risk functional constructed on the training sample from which the corresponding elements are excluded. As in analogous situations above, minimization of the functional (D-II.26) should be carried out using a heuristic procedure of successive minimization.

† Observe that such a definition of the structure is an *a priori* one only for the formulation of the problem of estimating values of functions where the matrix Φ is formed using all the $l + k$ vectors of the complete sample. Nevertheless we shall use the structure (D-II.25) here.

Above, when constructing the linear regression, the structure (D-II.25) was defined in accordance with the order in which the terms of the series (D-II.23) appear. Often, however, the order in (D-II.23) is determined not in accordance with the magnitude of the eigenvalues but in the course of constructing the regression. We now consider such a stepwise algorithm of regression construction. First one factor (the function $\chi_1(x)$) is selected by means of which the best approximation to empirical data is attained. To determine such a factor the problem of constructing a regression for one factor is solved n times (where n is the number of factors), and that factor is selected for which the value of the empirical risk is minimal. This factor is then fixed, and after that a second factor is selected using enumeration of the remaining $n - 1$ factors, such that the linear function constructed for these two factors yields the smallest value for the empirical risk. The second factor obtained is then fixed, and a third factor is selected, and so on.

As a final solution—using this procedure of ordering with respect to factors—we choose the function which yields the minimum over α and p for the functional

$$R(\alpha, p) = \left[\frac{I_{\text{emp}}(\alpha)}{1 - \sqrt{\dfrac{p\left(\ln \dfrac{l}{p} + 1\right) + \ln p\left(n - \dfrac{p-1}{2}\right) - \ln \eta}{l}}} \right]_{\infty}$$

This algorithm is basic for all the algorithms for the estimation of functional values at given points described in Chapter 10.

Bibliographical Remarks

Chapter 1

The problem of minimizing the expected risk on the basis of empirical data is one of the basic problems of applied mathematical analysis. It has been studied by many authors: L. LeCam [95, 96], P. Huber [87, 90], Ya. Z. Tsypkin [67, 68], V. N. Vapnik [7–15], A. Ya. Chervonenkis [11–15], and others.

In this book a special class of problems of minimizing the expected risk is considered—the class of problems of estimating dependences, which contains the problems of pattern recognition, regression estimation, and interpreting results of indirect experiments.

The theory of pattern recognition was initiated in the late fifties. In the sixties and seventies monographs written by the following authors (among others) were devoted to this subject: M. A. Aizerman *et al.* [2], Ya. Z. Tsypkin [66, 67], V. N. Vapnik and A. Ya. Chervonenkis [12], N. G. Zagoruiko [19], V. A. Kovalevskiĭ [25], K. S. Fu [64], N. J. Nilsson [43], V. N. Fomin [62], and K. Fukunaga [64a].

The problem of estimating regression can be traced to Gauss's time. A voluminous literature is devoted to this subject, in particular the classical texts of C. R. Rao [49] Yu. V. Linnik [34], and M. G. Kendall and A. Stuart [24].

Finally, the problem of interpreting results of indirect experiments reduces to solving operator equations which form ill-posed problems. The theory of ill-posed problems received wide attention during the period of the 1950s up to the 1970s (see bibliography in [56]). Among these contributions we mention the monograph by A. N. Tikhonov and V. Ya. Arsenin [56]. The Appendix to Chapter 1 is based on this work.

In this book a special class of stochastic ill-posed problems is singled out —those dealing with interpreting results of indicrect experiments.

Chapter 2

Applications of methods of stochastic approximation for solving problems of minimizing the expected risk based on large sample data can be found in the works of Ya. Z. Tsypkin [66, 67] and M. A. Aizerman *et al.* [2]. In these references, along with conditions for the convergence of procedures of the stochastic-approximation type, specific applications to the problem of pattern recognition and regression estimation are considered. Mathematical problems of the theory of stochastic approximation are discussed in the monograph by M. B. Nevel'son and P. Z. Has'minskiĭ [42].

When minimizing a functional of the expected risk using a limited set of empirical data, two approaches are distinguished: the classical approach based on methods of parametric statistics and the approach based on minimization of the empirical risk.

Methods of parametric statistics were developed in the twenties through the forties and are associated with the names of the famous statisticians R. A. Fisher, K. Pearson, and H. Cramèr, among others. At the present time, methods of parametric statistics are a working tool for solving numerous problems. They are discussed in all texts on statistics. See, for example, S. S. Wilks [58] and M. G. Kendall and A. Stuart [24].

The problem of applicability of methods of minimizing the empirical risk to determining the minimum of the expected risk arose more recently. In 1953 L. LeCam [95] showed that for certain classes of loss functions the method of minimizing empirical risk as the sample size increases determines a function minimizing the expected risk. In this paper [95] LeCam, for the first time, connected the problem of risk minimization with the conditions for uniform convergence of the means to the mathematical expectations and obtained conditions for uniform convergence in the case of certain types of loss functions. In 1968 P. Huber [87] showed that the method of minimizing the empirical risk is applicable to a wider class of loss functions. However, both LeCam's and Huber's papers investigate asymptotic possibilities of the method.

In 1971 V. N. Vapnik and A. Ya. Chervonenkis [11] obtained necessary and sufficient conditions for uniform convergence of frequencies of events to their probabilities and derived bounds on the rate of this convergence. Based on these bounds, it was possible to establish the applicability of the method of minimizing the empirical risk to problems of pattern recognition based on samples of limited size. Later in 1974 this result was extended to problems of estimating dependences of a more general nature (Vapnik and Chervonenkis [13]).

Chapter 3

Numerous papers are devoted to the problem of estimating densities specified up to a finite number of parameters [49, 34, 24]. However, up until recently all the problems in this direction were actually concerned with estimating unknown parameters of the density rather than estimating the density function. Only in 1965 did D. G. Keehn [93] obtain a Bayesian estimator of the density of a normal distribution (presented in Section 7), which as it turned out does not belong to the class of normal distributions.

In 1969 P. Ya. Lumel'skiĭ and P. N. Sapozhnikov obtained the best unbiased estimator for a density of a univariate distribution [36]. (This result is presented in Section 10.) Earlier A. N. Kolmogorov and V. M. Tihomirov [27] had obtained the best unbiased estimator for a univariate density.

As far as the problem of estimating parameters is concerned, the basic results were obtained by R. A. Fisher [82]. These results constitute the foundations of methods of parametric analysis.

The problem of discriminant analysis is essentially based on constructing a linear discriminant function. It was initially formulated by R. A. Fisher [82], who suggested minimizing the functional presented in Section 2. In 1966 the problem of constructing a linear discriminant function for normal distributions was solved by T. W. Anderson and R. R. Bahadur [71].

Other investigations in this area are associated with an attempt to write down a functional whose minimization leads to constructing a linear discriminant function applicable to more general distributions than normal ones. Initially Fisher's functional was used for this purpose, and later other functionals were considered. A detailed survey of the literature on discriminant analysis is presented in [60].

The case of independently distributed discrete features was also considered in discriminant analysis.

In 1952 A. M. Uttley constructed a discriminant automaton whose algorithm in essence differs only slightly from the modern discriminant automata constructed in accordance with the assumption of independence of discrete features [105].

Chapter 4

The idea of a robust method of estimation of a location parameter in a given class of densities is due to P. Huber. In 1967 he obtained a robust method of estimating a location parameter in the class of densities defined by a mixture [88]. (Huber's result is presented in Section 8.)

Later, other authors derived robust procedures for estimating location parameters for various classes of functions. In particular robust estimators were obtained in a class of densities concentrated basically on an interval,

in a class of densities close to normal ones, and so on. A detailed survey of available procedures for robust estimation is given in B. T. Polyak and Ya. Z. Tsypkin's paper [45].

Applications of procedures for robust estimation of location parameters to estimating regression parameters are also due to Polyak and Tsypkin [46]. Using various prototype examples, these authors have demonstrated the advantage of robust estimation of regression parameters in the case of samples of a limited size.

Chapter 5

Parameters estimation is the traditional method for solving problems of regression estimation. The central theme of the theory of estimating regression parameters based on samples of a limited size is the investigation of the least-squares method, establishing its optimal properties (the theorem on normal regression, the Gauss–Markov theorem).

These theorems establish the optimality of the least squares method in a class of certain methods. It was however often assumed that the least-squares method not only is the best method for estimating parameters in a given limited class, but is a "good" method in general (in a rather wide class of methods).

In 1956 C. Stein [103] unexpectedly produced an example which shows that the best estimator of the mean of a multivariate normal distribution with a known covariance matrix $\sigma^2 I$ (where σ^2 is a known number and I is a unit matrix) differs from the vector of realizations (i.e., is not the one obtained using the least-squares method).

In 1961 James and Stein [91] discovered a method of estimating the mean of a multivariate normal distribution with unknown σ^2 of the covariance matrix $\sigma^2 I$ which is uniformly better than the one obtained by means of realizations. Finally in 1970 A. Baranchik [73] found a whole class of such estimators. This class of estimators is presented in this book for obtaining estimators of parameters of a normal regression which are uniformly better than the least-squares estimators. The method for constructing estimators of regression parameters utilizing James–Stein–Baranchik estimators, presented in Section 3, was obtained using P. K. Bhattacharya's theorem [75].

Stein's example demonstrated that the supposition that unbiased methods of estimation always contain "good" estimators is unfounded. (Even for the simplest situation estimation methods uniformly better than the classical ones are available.)

The theory of constructing the best linear method of estimation is due to V. A. Koshcheev [31]. The theory allows us—utilizing some prior information—to obtain linear estimators which are superior to those following from the least-squares method.

However, the question whether there exists a method of estimating regression parameters which is better than the least-squares method in the case where no additional prior information is utilized is still open. This problem is connected with constructing a method of estimating the mean which is uniformly better than the empirical mean for random vectors which are realizations of a not necessarily normal distribution. In other words, the problem is reduced to the determination of estimators of the Stein type which are invariant relative to probability density functions. Such estimators are possible (see, e.g., the paper by J. O. Berger [74]).

Chapter 6

The problem of uniform convergence of frequencies of occurrence of events to their probabilities was first studied in the papers of V. I. Glivenko [85] and F. P. Cantelli [92]. They showed in 1933 that a uniform convergence of an empirical cumulative distribution function to the true one (i.e. a uniform convergence of frequencies to probabilities for a special class of events) is valid. In the same year, A. N. Kolmogorov [94] obtained an asymptotic bound on the rate of convergence, which was later refined by N. V. Smirnov [53].

Justification for the applicability of the method of minimizing the empirical risk to problems of pattern recognition is connected with the determination of the conditions for uniform convergence of frequencies to probabilities in arbitrary classes of events. In 1971 V. N. Vapnik and A. Ya. Chervonenkis [11] obtained necessary and sufficient conditions for uniform convergence of frequencies of occurrences of events to their probabilities for an arbitrary system of events and obtained bounds on the rate of this convergence. Some of these results (the sufficient conditions) are contained in the Appendix to Chapter 6. Necessary and sufficient conditions are covered in the Appendix to Chapter 7.

Chapter 7

The content of Chapter 7 is a direct generalization of the results obtained for bounds on the rate of uniform relative deviation of frequencies from their probabilities to the case of bounds on the rate of uniform relative deviation of the means from their mathematical expectations. These were obtained by V. N. Vapnik and A. Ya. Chervonenkis in 1974 [13] for the case when the ratio of the means of order $p \geq 2$ of positive random variables to their means is uniformly bounded. The case when the ratio of the means of order $1 < p \leq 2$ to their means is uniformly bounded is examined here for the first time.

In 1981 V. N. Vapnik and A. Ya. Chervonenkis obtained necessary and sufficient conditions for uniform convergence of means to their mathematical expectations for uniformly bounded functions [15]. This theory is incorporated in the present edition of this book (Appendix to Chapter 7).

Chapter 8

The method of structural minimization of risk was formulated for solving pattern recognition problems in the monograph by V. N. Vapnik and A. Ya. Chervonenkis [12]. However, in essence, when constructing algorithms for minimizing risks the method is utilized each time when the method of minimizing empirical risk yields unreasonable results (for example, when estimating polynomial regressions).

A two-stage selection procedure (of an element of the structure and of the best function belonging to a given element) is contained in all heuristic algorithms whose purpose is to obtain a solution which is better than the one that follows from the standard methodology of minimizing the empirical risk (cf. for example, the papers by I. Sh. Pinsker [44] and A. G. Ivahnenko *et al.* [22]).

In this book two ideas are used for the selection criterion of an element of a structure: an estimate for the moving-control procedure and a uniform bound on the value of the expected risk based on the values of the empirical risk. Estimation of the expected risk follows from the theory of uniform convergence. For the moving-control procedure, it was A. L. Luntz and V. L. Brailovskiĭ [37] who established the unbiasedness of the estimator in 1969. In Chapter 8 an equivalent representation of the moving-control estimator for regressions is presented. The representation allows us to substantially reduce the amount of calculation.

The theory of $J_p(x_1, y_1; \ldots; x_l, y_l)$ estimators is presented here for the first time. Also the sample selection is considered here for the first time.

Chapter 9

The idea of applying the method of structural minimization of the risk to ill-posed problems of measurements was implemented for the first time in the paper by V. N. Vapnik and A. I. Mihal'skiĭ [8]. However, various (heuristic) devices were used earlier which allowed the choice of an appropriate form for approximation (cf. for example, the papers by L. A. Vaĭnsteĭn [6] and L. P. Grabar' [18]).

In 1975 V. N. Vapnik and A. Ya. Chervonenkis established the existence of convergence—as the number of empirical data increases—of a sequence of solutions obtained using the method of structural minimization of the risk to the desired solution under the condition that the solution is sought in the form of an expansion in terms of a special system of functions; if, however, the solution is sought in the form of an expansion in terms of polynomials, then such convergence may not occur.

In 1974 A. I. Mihal'skiĭ showed that for certain classes of operator equations there exists convergence—as the sample size increases—of solutions determined using the the method of structural risk minimization to the desired function provided the solution is sought in a class of splines. He also developed a methodology for constructing splines with a given number of conjugates minimizing the empirical risk [38].

The idea of estimating the probability density by viewing it as a solution to an ill-posed problem of numerical differentiation was implemented in V. N. Vapnik and A. R. Stefanyuk's paper [10]. In this book a generalization to the stochastic case of A. N. Tihonov's theorem is presented in the Appendix to Chapter 9. A bound on the rate of convergence of approximations to a smooth density derived by means of the regularization method was obtained by A. R. Stefanyuk [53a].

Chapter 10

The problem of estimating values of a function at given points was considered for the first time in V. N. Vapnik and A. Ya. Chervonenkis's monograph [12] for the case of indicator functions. In V. N. Vapnik and A. M. Sterin's paper [9] various structures on equivalence classes of indicator functions were studied.

Methods for estimating of an arbitrary function at given points are considered here for the first time. Also new is the investigation presented in this chapter of selecting a complete sample and finding the best point in a given set.

Addenda I and II

A library of programs for the method of generalized portraits has been developed by T. G. Glazkova and A. A. Zuravel'. Algorithms for estimating values of indicator functions at given points have been studied and implemented by A. M. Sterin.

Algorithms for regression estimation have been compiled by T. G. Glaskova, V. A. Koshcheev, and A. I. Mihal'skiĭ.

Algorithms for interpreting ill-posed problems of measurements have been devised by A. I. Mihal'skiĭ.

Bibliography

Translator's remark. Whenever possible the English version of the paper or book cited is presented. However, the order of entries follows the Russian edition. This has been done to minimize the possibility of printing errors. Since all the references to the literature in the text are specified by the ordinal number, this ordering should cause no confusion.

1. Aivazyan, S. A., Bezhaeva, Z. I., and Staroverov, O. V.: *Classification of Multivariate Observations.* Moscow: Statistika 1974.
2. Aizerman, M. A., Braverman, E. M., and Rozonoer, L. I.: *Method of Potential Function in the Theory of Machine Learning.* Moscow: Nauka 1970.
3. Vapnik, V. N. (Ed.): *Algorithms for Pattern Recognition.* Moscow: Sovetskoe Radio 1973.
4. Ahlberg, J. H., Nilson, E. N., and Walsh, J. L.: *The Theory of Splines and Their Applications.* New York: Academic Press 1967.
5. Anderson, T. W.: *An Introduction to Multivariate Statistical Analysis.* New York: Wiley 1958.
6. Vaĭnsteĭn, L. A.: On the numerical solution of integral equations of the first type utilizing prior information on the estimated function, *Dokl. Akad. Nauk SSSR*, **204**, No. 6, 1331–1334 (1972) (omitted from translation in *Soviet Math.*, **13**, No. 3).
7. Vapnik, V. N. (Ed.): *Machines for Pattern Recognition—Algorithms for Pattern Recognition.* Moscow: Sovetskoe Radio 1973.
8. Vapnik, V. N., and Mihal'skiĭ, A. I.: On a search for dependences using the method of structural risk minimization, *Avtomatika and Telemehanika*, No. 10 (1974).
9. Vapnik, V. N., and Sterin, A. M.: On structural minimization of the overall risk in a problem of pattern recognition, *ibid.*, No. 10 (1977).
10. Vapnik, V. N., and Stefanyuk, A. R.: Nonparametric methods for estimating probability densities, *ibid.* No. 8 (1978).
11. Vapnik, V. N., and Chervonenkis, A. Ya.: On the uniform convergence of relative frequencies of events to their probabilities, *Theor. Prob. Appl.*, **16**, 264–280 (1971).
12. Vapnik, V. N., and Chervonenkis, A. Ya.: *The Theory of Pattern Recognition.* Moscow: Nauka 1974.

13. Vapnik, V. N., and Chervonenkis, A. Ya.: On a method of structural risk minimization, *Avtomatika and Telemehanika*, No. 8, 9 (1974).
14. Vapnik, V. N. and Chervonenkis, A. Ya.: On asymptotic properties of the method of structural risk minimization, *ibid.*, No. 12 (1975).
15. Vapnik, V. N., and Chervonenkis, A. Ya.: Necessary and sufficient conditions for uniform convergence of means to mathematical expectations, *Teor. Veroyatn. i Primen.* **26**, No. 3 (1981).
16. Vitushkin, A. G.: *Estimation of the Complexity of the Tabulation Problem.* Moscow: Fizmatgiz 1959.
17. Gnedenko, B. V.: *The Theory of Probability* (transl. by B. D. Seckler). New York: Chelsea 1962.
18. Grabar', L. P.: An application of Chebyshev polynomials orthonormalized on a system of equidistant points for solving integral equations of the first kind, *Soviet Math.*, **8**, 164–167 (1967).
19. Zagoruĭko, N. G.: *Methods of Recognition and Their Applications.* Moscow: Sovetskoe Radio 1972.
20. Ivanov, V. K.: On linear problems which are not well posed, *Soviet Math.* **3**, 981–983 (1962).
21. Ivanov, V. K.: On ill-posed problems, *Mat. Sbornik*, **61**, No. 2 (1963) [Russian].
22. Ivahnenko, A. G., Zaichencko, Yu. P., and Dmitrov, V. D.: *Decision Making Based on Self-Organization.* Moscow: Sovetskoe Radio 1976.
23. Kagan, A. M., and Shalaevskii, O. V.: Admissibility of estimators of least squares —a characterizing property of the normal law, *Mathematical Notes*, **6**, No. 1 (1969).
24. Kendall, M. G., and Stuart, A.: *Advanced Theory of Statistics, Vol. 2, Inference and Relationship*, 3rd ed., London: Griffin 1977.
25. Kovalevskiĭ, V. A.: *Methods of Optimal Solutions in Recognition of Images.* Moscow: Nauka 1976.
26. Kolmogorov, A. N.: Unbiased estimates, *Izv. A.N. SSSR, Math.*, No. 4, 303–326 (1950) [Russian].
27. Kolmogorov, A. N., and Tihomirov, V. M.: ε-entropy and ε-capacity in functional spaces, *Uspehi Mat. Nauk*, No. 2 (1959).
28. Kolmogorov, A. N., and Fomin, S. V.: *Introductory Real Analysis* (rev. English ed., transl. and ed. by R. A. Silverman). Englewood Cliffs, N.J.: Prentice Hall 1970.
29. Korbut, A. A., and Finkel'shteĭn, Yu. Yu.: *Discrete Programming*; Moscow: Nauka (D. B. Yudin, Ed.) 1969.
30. Korovkin, P. P.: *Linear Operators and Approximation Theory.* Delhi: Hindustan Publ. Corp. 1960.
31. Koshcheev, V. A.: Methods of taking prior information into account in linear estimation of parameters, in *Statistical Methods of Control Theory.* Moscow: Nauka 1978.
32. Lavrent'ev, M. M.: *On Some Ill-Posed Problems of Mathematical Physics.* Publication of the Siberian Branch of the Academy of Sciences of the USSR, 1962.
33. LeCam, L.: On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, in *Univ. of Calif. Publ. Statistics*, Vol. 1, pp. 277–329. Berkeley: Univ. of Calif. Press, 1953.
34. Linnik, Yu. V.: *Method of Least Squares* (trans. by R. Elandt). New York: Pergamon Press 1961.
35. Linnik, Ju. V.: *Statistical Problems with Nuisance Parameters* (transl. from Russian by Scripta Technica). Providence: Amer. Math. Soc. 1968.
36. Lumel'skiĭ, P. Ya., and Sapozhnikov, P. N.: Unbiased estimates of density functions, *Theor. Prob. Appl.*, **14**, 357–365 (1969).

37. Luntz, A. L., and Brailovskiĭ, V. L.: On estimation of characters obtained in statistical procedures of recognition, *Izv. Akad. Nauk SSSR, Ser. Tehnicheskaya Kibernetika*, No. 3 (1969).
38. Mihal'skiĭ, A. I.: A method of averaged splines in the problem of approximating dependences based on empirical data, *Avtomatika and Telemehanika*, No. 3 (1974).
38a. Mihal'skiĭ, A. I.: Estimation of statistical dependence by means of averaged splines, *Vychislit. Matem. i Mat. Fizika*, No. 5, 1107–1117 (1979).
39. Morozov, V. A.: The error principle in the solution of incompatible equations by Tikhonov regularization, *USSR Computational Mathematics and Mathematical Physics*, **13**, No. 5, 1–16 (1973).
40. Morozov, V. A.: Calculation of the lower bounds of functionals from approximate information, *ibid.* **13**, No. 4, 275–281 (1973).
41. Nadaraya, E. A.: On nonparametric estimates of density functions and regression curves, *Theor. Prob. Appl.*, **10**, 186–190 (1965).
42. Nevel'son, M. B., and Has'minskiĭ, R. Z.: *Stochastic Approximation and Recursive Estimation*: Moscow: Nauka 1972.
43. Nilsson, N. J.: *Learning Machines: Foundations of Trainable Pattern-Classifying Systems.* New York: McGraw-Hill 1965.
44. Pinsker, I. Sh.: A selection of a structure and computation of parameters of a decision rule under a limited sample, in *Modelling and Automatic Analysis of Electrocardiograms.* Moscow: Nauka 1973.
45. Polyak, B. T., and Tsypkin, Ya. Z.: Pseudogradient algorithms of adaptation and learning, *Avtomatika i Telemehanika*, No. 3 (1973).
46. Polyak, B. T., and Tsypkin, Ya. Z.: Error-stable identification, in *Proceedings of the IV Symposium of IFAK on Identification*, Part I. Tbilisi: Metsniereba 1976.
47. Pugachev, V. S.: Statistical theory of learning automatic systems, *Tehnicheskaya Kibernetika*, No. 6 (1967).
48. Pugachev, V. S.: Statistical problems of the theory of pattern recognition, in *Proceedings of the III All Union Conference on Automatic Control.* Moscow: Nauka 1967.
49. Rao, C. R.: *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley 1973.
50. Rastrigin, L. A., and Erenshteĭn, R. H.: Decision making by means of a collective of decision rules in problems of pattern recognition, *Avtomatika i Telemehanika*, No. 9 (1975).
51. Raudis, Sh. Yu.: Limitations of selection in problems of classification, in *Statistical Problems of Control*, No. 18. Vil'nyus: Institute of Physics and Mathematics, Akad. Nauk Litovsk. SSR 1977.
52. Ryžik, I. M., and Gradshteĭn, I. S.: *Tables of Integrals, Sums, Series, and Products.* Moscow: Gostehizdat 1956.
53. Smirnov, N. V.: *Theory of Probability and Mathematical Statistics* (*Selected Works*). Moscow: Nauka 1970.
53a. Stefanyuk, A. R.: On the rate of convergence of a class of estimators for probability density, *Avtomatika i Telemehanika*, No. 11 (1979).
54. Tikhonov, A. N.: Solution of incorrectly formulated problems and the regularization method, *Soviet Math.*, **4**, 1035–1038 (1963).
55. Tikhonov, A. N.: Regularization of incorrectly posed problems, *Soviet Math.*, **4**, 1624–1627 (1963).
56. Tikhonov, A. N., and Arsenin, V. Ya.: *Solutions of Ill-posed Problems.* Washington: Winston 1977.
57. Turchin, V. F., Kozlov, V. N., and Malkevich, M. S.: Utilization of methods of mathematical statistics for solving ill-posed problems, *Uspehi Fiz. Nauk*, **102**, No. 3 (1970).

58. Wilks, S. S.: *Mathematical Statistics.* New York: Wiley 1962.
59. Wilkinson, J. H.: *The Algebraic Eigenvalue Problem.* Oxford: Clarendon Press 1965.
60. Urbah, V. Yu.: Discriminant analysis: basic ideas and applications, in *Statistical Methods of Classification*, No. 1. Moscow: State University Press 1969.
61. Feller, W.: *An Introduction to Probability Theory and Its Applications*, Vol. 1, 2nd ed. New York: Wiley 1957.
62. Fomin, V. N.: *Mathematical Theory of (Learning) Recognition Systems*; Leningrad: State University Press 1976.
63. Forsythe, G. E., and Moler, C. B.: *Computer Solution of Linear Algebraic Systems.* Englewood Cliffs, N.J.: Prentice-Hall 1967.
64. Fu, K. S.: *Syntactic Methods in Pattern Recognition.* New York: Academic Press 1974.
64a. Fukanaga, K.: *Introduction to Statistical Pattern Recognition*; New York: Academic Press 1972.
65. Hadley, G.: *Non-linear and Dynamic Programming.* Reading, Mass.: Addison-Wesley 1964.
65a. Khinchin, A. Ya.: On basic theorems of information theory, *Uspehi Math. Nauk*, **XI**, 17–75 (1956) [Russian].
66. Tsypkin, Ya. Z.: *Adaptation and Learning in Automatic Systems.* New York: Academic Press 1971.
67. Tsypkin, Ya. Z.: *Foundations of the Theory of Learning Systems.* New York: Academic Press 1973.
68. Čencov, N. N.: Evaluation of an unknown distribution density from observations, *Soviet Math.*, **4**, No. 6, 1559–1562 (1963).
69. Jakubovič, V. A.: Some general theoretical principles for constructing learning recognition systems, in *Vychislit. Tekhnika i Voprosy Program.* Leningrad: University Press 1965.
70. Jakubovič, V. A.: Finitely convergent recursive algorithms for the solution of the system of inequalities, *Soviet Math.*, **7**, No. 1, 300–304 (1966).
71. Anderson, T. W., and Bahadur, R. R.: Classification into two multivariate normal distributions with different covariance matrices, *Ann. Math. Stat.*, **33**, No. 2 (1966).
72. Andrews, H.: *Introduction to Mathematical Techniques in Pattern Recognition.* New York: Wiley 1972.
73. Baranchik, A. J.: A family of minimax estimators of the mean of a multivariate normal distribution, *Ann. Math. Stat.*, **41**, No. 2 (1970).
74. Berger, J. O.: Minimax estimation of location vectors of a wide class of densities, *Ann. Stat.*, **3**, No. 6 (1975).
75. Bhattacharya, P. K.: Estimating the mean of a multivariate normal population with general quadratic loss function, *Ann. Math. Stat.*, **37**, No. 18 (1966).
75a. Bowker, A. H., and Sitgreaves, R.: An asymptotic expansion to the distribution function of the *W*-classification statistic, in *Studies in Item Analysis and Prediction* (H. Solomon, Ed.). Stanford, CA: Stanford University Press 1961.
76. Cacoullos, T.: Estimation of a multivariate density, *Inst. Stat. Math. Tokyo*, **18**, No. 2 (1966).
77. Chen, C. H.: *Statistical Pattern Recognition.* New York: Hayden 1973.
78. Devroye, L. P., and Wagner, T. J.: A distribution-free performance bound in error estimation, *IEEE Trans. Info. Theory*, **IT-22**, No. 5 (1976).
79. Dvoretsky, A., On stochastic approximation, in *Proceedings of the III Berkeley Symposium on Mathematical Statistics and Probability*, Vol. **1** 1956.
80. Dancel, J. W.: The conjugate gradient method for linear and nonlinear operator equations, *SIAM J. Number. Anal.*, **4**, No. 1 (1967).
80a. Golin, A., Kline, I., Sofina, Z. P., and Syrkin, A. B. (Eds.), *Experimental Evaluation of Antitumor Drugs in the USA and USSR and Clinical Correlations*; Bethseda 1980.

81. Fix, I. R., and Hodges, J. L.: Discriminatory analysis; nonparametric discrimination: consistency properties; *Report 4 of the USAF School of Aviation Medicine.* Randolph Field. Texas 1952.
82. Fisher, R. A.: *Contributions to Mathematical Statistics.* New York: Wiley 1952.
83. Fraser, D. A. S.: *Nonparametric Method in Statistics.* New York: Wiley 1957.
84. Fukunaga, K.: *Introduction to Statistical Pattern Recognition.* New York: Academic Press 1972.
85. Glivenko, V. I.: Sulla determinatione empirica di una legge di probabilita, *Giornale dell' Istituto Italiano degli Attuari*, **4** (1933).
86. Hestenes, M. R., and Stiefel, E.: Method of conjugate gradients for solving linear systems, *J. Res. Mat., Bur. Stand.*, **49**, No. 6 (1952).
87. Huber, P.: The behavior of maximum likelihood estimates under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1 (1967).
88. Huber, P.: Robust estimation of the location parameter, *Ann. Math. Stat.*, **35**, No. 1 (1964).
89. Huber, P.: Robust statistics: a review, *Ann. Math. Stat.*, **43**, No. 4 (1972).
90. Huber, P.: Robust regression: asymptotics, conjectures and Monte Carlo, *Ann. Stat.*, **1**, No. 5 (1973).
91. James, W., and Stein, C.: Estimation with quadratic loss, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Univ. of Calif. Press, 1961.
92. Cantelli, F. P.: Sulla determinazione empirica della leggi di probabilita, *Giornale dell'Istituto Italiano degli Attuari*, **4** (1933).
93. Keehn, D. G.: Note on learning for Gaussian properties, *IEEE Trans. Info. Theory*, **IT-11**, No. 1 (1965).
94. Kolmogorov, A. N.: Sulla determinatione empirica di una legge di distributione, *Giornale dell'Istituto Italiano degli Attuari*, **4** (1933).
95. LeCam, L.: On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, *Univ. of Calif. Public. Statist.*, **11** (1953).
96. LeCam, L.: On asymptotic theory of estimation and testing hypotheses, in *Proceedings of the III Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. 1956.
97. Lawler, E. L., and Wood, D. E.: Branch-and-bound method, survey, *Oper. Res.*, **14**, No. 4, 699–717 (1966).
98. Meisel, W. S.: *Computer-Oriented Approaches to Pattern Recognition.* New York: Academic Press 1972.
99. Parzen, E.: On the estimation of a probability function and mode, *Ann. Math. Stat.*, **33**, No. 3 (1962).
100. Reiss, R. D.: Consistency of a certain class of density function, *Metrika*, **22**, No. 4 (1975).
101. Sacks, J., and Ylvisaker, D.: A note on Huber's robust estimation of a location parameter, *Ann. Math. Stat.*, **43**, No. 4 (1972).
102. Sclove, S. L.: Improved estimators for coefficients in linear regression, *JASA*, **63**, No. 322 (1968).
103. Stein, C.: Inadmissibility of usual estimator for mean of a multivariate normal distribution, in *Proceedings of the Third Berkeley Symposium in Mathematical Statistics and Probability*, Vol. 1. Univ. of Calif. Press 1956.
104. Wald, A.: Note on the consistency of the maximum likelihood estimate, *Ann. Math. Stat.*, **20** (1949).
105. Uttley, A. M.: A theory of the mechanism of learning on the computation of conditional probabilities, in *Proceeding of the First Congress on International Cybernetics.* Namur 1956.

Literature Added in the English Edition

106. Allen, D.: The relationship between variable selection and data augmentation and a method of prediction, *Technometrics*, **16**, No. 1, 125–127 (1974).
107. Belsley, D. A., Kuh, E., and Welsch, R. E.: *Regression Diagnostics*. New York: Wiley 1980.
108. Berger, J.: *Statistical Decision Theory: Foundations, Concepts and Methods*. New York: Springer 1980.
109. Bickel, P., and M. Rosenblatt.: On some global measures of the deviations of density function estimates, *Ann. Statist.*, **1**, 1071–1095 (1973).
110. Brown, L.: Admissible estimator, recurrent defusions and insoluble boundary value problems, *Ann. Math. Statist.*, **42**, 855–903 (1971).
111. Dudley, R.: Central limit theorems for empirical processes, *Ann. Probability*, **6**, 899–929 (1978).
112. Grenander, U.: *Abstract Inference*. New York: Wiley 1981.
113. Hasminsky, R. Z.: A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Probability and Appl.*, **23**, 794–798 (1978).
114. Hoerl, A. E., and Kennard, R. W.: Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67 (1970).
115. Huber, P. J.: *Robust Statistics*: New York: Wiley 1981.
116. Stone, M.: Cross-validatory choice and assessment of statistical predictions, *JRSS.*, **B36**, 111–147 (1974).
117. Tapia, R. A., and Thompson, J. R.: *Nonparametric Probability Density Estimation*. Baltimore: Johns Hopkins Univ. Press 1978.
118. Wegman, E. J.: Nonparametric probability density estimation, I. A summary of available methods. *Technometrics*, **14**, 533–546 (1972).
119. Wenocur, R. S., and Dudley, R. M.: Some special Vapnik–Chervonenkis classes, *Discrete Mathematics*, **33**, 313–318 (1981).
120. Wertz, W., and Schneider, B.: Statistical density estimation: A Bibliography, *International Statist. Review*, **47**, 155–175 (1979).

Index

(The terms which appear prominently in the headings of the chapters and sections are generally not included in the Index.)

Vladimir Vapnik

Empirical Inference Science

Afterword of 2006

With 6 Illustrations

Springer

To the students of my students in memory of my violin teacher Ilia Shtein and PhD advisor Alexander Lerner, who taught me several important things that are very difficult to learn from books.

PREFACE

Twenty-five years have passed since the publication of the Russian version of the book *Estimation of Dependencies Based on Empirical Data* (*EDBED* for short). Twenty-five years is a long period of time. During these years many things have happened. Looking back, one can see how rapidly life and technology have changed, and how slow and difficult it is to change the theoretical foundation of the technology and its philosophy.

I pursued two goals writing this Afterword: to update the technical results presented in *EDBED* (the easy goal) and to describe a general picture of how the new ideas developed over these years (a much more difficult goal).

The picture which I would like to present is a very personal (and therefore very biased) account of the development of one particular branch of science, *Empirical Inference Science*.

Such accounts usually are not included in the content of technical publications. I have followed this rule in all of my previous books. But this time I would like to violate it for the following reasons. First of all, for me *EDBED* is the important milestone in the development of empirical inference theory and I would like to explain why. Second, during these years, there were a lot of discussions between supporters of the new paradigm (now it is called the VC theory[1]) and the old one (classical statistics). Being involved in these discussions from the very beginning I feel that it is my obligation to describe the main events.

The story related to the book, which I would like to tell, is the story of how it is difficult to overcome existing prejudices (both scientific and social), and how one should be careful when evaluating and interpreting new technical concepts.

This story can be split into three parts that reflect three main ideas in the development of empirical inference science: from the pure technical (mathematical) elements of the theory to a new paradigm in the philosophy of generalization.

[1] VC theory is an abbreviation for Vapnik–Chervonenkis theory. This name for the corresponding theory appeared in the 1990s after *EDBED* was published.

The first part of the story, which describes the main technical concepts behind the new mathematical and philosophical paradigm, can be titled

Realism and Instrumentalism: Classical Statistics and VC Theory

In this part I try to explain why between 1960 and 1980 a new approach to empirical inference science was developed in contrast to the existing classical statistics approach developed between 1930 and 1960.

The second part of the story is devoted to the rational justification of the new ideas of inference developed between 1980 and 2000. It can be titled

Falsifiability and Parsimony: VC Dimension and the Number of Entities

It describes why the concept of VC falsifiability is more relevant for predictive generalization problems than the classical concept of parsimony that is used both in classical philosophy and statistics.

The third part of the story, which started in the 2000s can be titled

Noninductive Methods of Inference: Direct Inference Instead of Generalization

It deals with the ongoing attempts to construct new predictive methods (direct inference) based on the new philosophy that is relevant to a complex world, in contrast to the existing methods that were developed based on the classical philosophy introduced for a simple world.

I wrote this Afterword with my students' students in mind, those who just began their careers in science. To be successful they should learn something very important that is not easy to find in academic publications.

In particular they should see the big picture: what is going on in the development of this science and in closely related branches of science in general (not only about some technical details). They also should know about the existence of very intense paradigm wars. They should understand that the remark of Cicero, "Among all features describing genius the most important is inner professional honesty", is not about ethics but about an intellectual imperative. They should know that Albert Einstein's observation about everyday scientific life that "Great spirits have always encountered violent opposition from mediocre minds," is still true. Knowledge of these things can help them to make the right decisions and avoid the wrong ones. Therefore I wrote a fourth part to this Afterword that can be titled

The Big Picture.

This, however, is an extremely difficult subject. That is why it is wise to avoid it in technical books, and risky to discuss it commenting on some more or less recent events in the development of the science.

Writing this Afterword was a difficult project for me and I was able to complete it in the way that it is written due to the strong support and help of my colleagues Mike Miller, David Waltz, Bernhard Schölkopf, Leon Bottou, and Ilya Muchnik.

I would like to express my deep gratitude to them.

Princeton, New Jersey, Vladimir Vapnik
November 2005

CONTENTS

Chapter 1

REALISM AND INSTRUMENTALISM: CLASSICAL STATISTICS AND VC THEORY (1960–1980)

1.1 THE BEGINNING

In the history of science two categories of intellectual giants played an important role:

(1) The giants that created the new models of nature such as Lavoisier, Dirac, and Pasteur;

(2) The giants that created a new vision, a new passion, and a new philosophy for dealing with nature such as Copernicus, Darwin, Tsiolkovsky, and Wiener.

In other words, there are giants who created new technical paradigms, and giants who created new conceptual (philosophical) paradigms. Among these, there are unique figures who did both, such as Isaac Newton and Albert Einstein.

Creating a new technical paradigm is always difficult. However, it is much more difficult to change a philosophical paradigm. To do this sometimes requires several generations of scientists.[1] Even now one can see the continuation of the old paradigm wars in articles discussing (in a negative way) the intellectual heritage of the great visionaries Charles Darwin, Albert Einstein, Norbert Wiener, and Isaac Newton.

My story is about attempts to shift one of the oldest philosophical paradigms related to the understanding of human intelligence. Let me start with the vision Wiener described in his book *Cybernetics*. The main message of this book was that there are no

[1] Fortunately scientific generations change reasonably fast, about every ten years.

big conceptual differences between solving intellectual problems by the brain or by a computer, and that it is possible to use computers to solve many intellectual problems.

Today every middle school student will agree with that (five scientific generations have passed since Wiener's time!). However, 50 years ago even such giants as Kolmogorov hesitated to accept this point of view.

1.1.1 The Perceptron

One of the first scientific realizations of Wiener's idea was a model of how the brain learns introduced by Rosenblatt. He created a computer program called the "Perceptron" and successfully checked it on the digit recognition problem. Very soon Novikoff proved that the Perceptron algorithm (inspired by pure neurophysiology) constructs a hyperplane in some high-dimensional feature space that separates the different categories of training vectors.

It should be mentioned that models of how the brain generalizes and different pattern recognition algorithms both existed at the time of the Perceptron. These algorithms demonstrated success in solving simple generalization problems (for example Selfridge's Pandemonium, or Steinbuch's Learning Matrix).

However, after Rosenblatt's Perceptron and Novikoff's theorem, it became clear that complex biological models can execute very simple mathematical ideas. Therefore it may be possible to understand the principles of the organization of the brain using abstract mathematical arguments applied to some general mathematical constructions (this was different from analysis of specific technical models suggested by physiologists).

1.1.2 Uniform Law of Large Numbers

The Novikoff theorem showed that a model of the brain described in standard physiological terms ("neurons," "reward and punishment," "stimulus") executes a very simple mathematical idea — it constructs a hyperplane that separates two different categories of data in some mathematical space. More generally, it minimizes in a given set of functions an empirical risk functional.

If it is true that by minimizing the empirical risk one can generalize, then one can construct more efficient minimization algorithms than the one that was used by the Perceptron. Therefore in the beginning of the 1960s many such algorithms were suggested. In particular Alexey Chervonenkis and I introduced the optimal separating hyperplane that was more efficient for solving practical problems than the Perceptron algorithm (especially for problems with a small sample size). In the 1990s this idea became a driving force for SVMs (we will discuss SVMs in Chapter 2, Section 2.3). However, just separation of the training data does not guarantee success on the test data. One can easily show that good separating of the training data is a necessary condition for the generalization. But what are the sufficient conditions?

This led to the main question of learning theory:

When does separation of the training data lead to generalization?

This question was not new. The problem, "How do humans generalize?" (What is the model of induction? Why is the rule that is correct for previous observations also correct for future observations?) was discussed in classical philosophy for many centuries. Now the same question — but posed for the simplest mathematical model of generalization, the pattern recognition problem — became the subject of interest.

In the beginning of the 1960s many researchers including Chervonenkis and I became involved in such discussions. We connected this question with the existence of uniform convergence of frequencies to their probabilities over a given set of events. To find the conditions that guarantee the generalization for the pattern recognition problem, it is sufficient to find the conditions for such convergence.

Very quickly we constructed a theory for uniform convergence over sets with a finite number of events (1964) and in four years we obtained the general answer, the necessary and sufficient conditions for uniform convergence for any (not necessarily finite) set of events. This path is described in *EDBED*.

What was not known at the time EDBED was written is that the uniform convergence describes not only sufficient conditions for generalization but also the necessary conditions:

Any algorithm that uses training data to choose a decision rule from the given admissible set of rules must satisfy it.

It took us another 20 years to prove this fact. In 1989 we proved the main theorem of VC theory that states:[2]

If the necessary and sufficient conditions for uniform convergence are not valid, that is, if the VC entropy over the number of observations does not converge to zero,

$$\frac{H_P^{\Lambda}(\ell)}{\ell} \longrightarrow c \neq 0,$$

then there exists a subspace X^* of the space R^n whose probability measure is equal to c,

$$P(X^*) = c,$$

such that almost any sample of vectors $x_1^*, \ldots, x_k^*$ of arbitrary size k from the subspace X^* can be separated in all 2^k possible ways by the functions from the admissible set of indicator functions $f(x, \alpha)$, $\alpha \in \Lambda$. (See also *EDBED*, Chapter 6 Section 7 for the definition of VC entropy).

This means that if uniform convergence does not take place then any algorithm that does not use additional prior information and picks up one function from the set of admissible functions cannot generalize.[3]

[2]Below for the sake of simplicity we formulate the theorem for the pattern recognition case (sets of indicator functions), but the theorem has been proven for any set of real-valued functions [121;140]. Also to simplify formulation of the theorem we used the concept of "two-sided uniform convergence" discussed in *EDBED* instead of "one-sided" introduced in [121].

[3]This, however, leaves an opportunity to use averaging algorithms that possess a priori information about the set of admissible functions. In other words VC theory does not intersect with Bayesian theory.

If, however, the conditions for uniform convergence are valid then (as shown in Chapter 6 of *EDBED*) for any fixed number of observations one can obtain a bound that defines the guaranteed risk of error for the chosen function.

Using classical statistics terminology the uniform convergence of the frequencies to their probability over a given set of events can be called the *uniform law of large numbers* over the corresponding set of events. (The convergence of frequencies to their corresponding probability for a fixed event (the Bernoulli law) is called the law of large numbers.)

Analysis of Bernoulli's law of large numbers has been the subject of intensive research since the 1930s. Also in the 1930s it was shown that for one particular set of events the uniform law of large numbers always holds. This fact is the Glivenco–Cantelli theorem. The corresponding bound on the rate of convergence forms Kolmogorov's bound. Classical statistics took advantage of these results (the Glivenco–Cantelli theorem and Kolmogorov's bound are regarded as the foundation of theoretical statistics).

However, to analyze the problem of generalization for pattern recognition, one should have an answer to the more general question:

> *What is the demarcation line that describes whether the uniform law of large numbers holds?*

The obtaining of the existence conditions for the uniform law of large numbers and the corresponding bound on the rate of convergence was the turning point in the studies of empirical inference.

This was not recognized immediately, however. It took at least two decades to understand this fact in full detail. We will talk about this in what follows.

1.2 Realism and Instrumentalism in Statistics and the Philosophy of Science

1.2.1 The Curse of Dimensionality and Classical Statistics

The results of successfully training a Perceptron (which constructed decision rules for the ten-class digit classification problem in 400-dimensional space, using 512 training examples) immediately attracted the attention of the theorists.

In classical statistics a problem analogous to the pattern recognition problem was considered by Ronald Fisher in the 1930s, the so-called problem of discriminant analysis. Fisher considered the following problem. One knows the generating model of data for each class, the density function defined up to a fixed number of parameters (usually Gaussian functions). The problem was: given the generative models (the model how the data are generated known up to values of its parameters) estimate the discriminative rule. The proposed solution was:

> *First*, using the data, estimate the parameters of the statistical laws and

Second, construct the optimal decision rule using the estimated parameters.

To estimate the densities, Fisher suggested the maximum likelihood method.

This scheme later was generalized for the case when the unknown density belonged to a nonparametric family. To estimate these generative models the methods of nonparametric statistics were used (see example in Chapter 2 Section 2.3.5). However, the main principle of finding the desired rule remained the same: first estimate the generative models of data and then use these models to find the discriminative rule.

This idea of constructing a decision rule after finding the generative models was later named the *generative model of induction.* This model is based on understanding of how the data are generated. In a wide philosophical sense an understanding of how data are generated reflects an understanding of the corresponding law of nature.

By the time the Perceptron was introduced, classical discriminant analysis based on Gaussian distribution functions had been studied in great detail. One of the important results obtained for a particular model (two Gaussian distributions with the same covariance matrix) is the introduction of a concept called the Mahalanobis distance. A bound on the classification accuracy of the constructed linear discriminant rule depends on a value of the Mahalanobis distance.

However, to construct this model using classical methods requires the estimation of about $0.5n^2$ parameters where n is the dimensionality of the space. Roughly speaking, to estimate one parameter of the model requires C examples. Therefore to solve the ten-digit recognition problem using the classical technique one needs $\approx 10(400)^2C$ examples. The Perceptron used only 512.

This shocked theorists. It looked as if the classical statistical approach failed to overcome the curse of dimensionality in a situation where a heuristic method that minimized the empirical loss easily overcame this curse.

Later the methods based on the idea of minimizing different type of empirical losses were called the *predictive (discriminative) models of induction*, in contrast to the classical *generative models.* In a wide philosophical sense predictive models do not necessarily connect prediction of an event with understanding of the law that governs the event; they are just looking for a function that explains the data best.[4]

The VC theory was constructed to justify the empirical risk minimization induction principle: according to VC theory the generalization bounds for the methods that minimize the empirical loss do not depend directly on the dimension of the space. Instead they depend on the so-called capacity factors of the admissible set of functions — the VC entropy, the Growth function, or the VC dimension — that can be much smaller than the dimensionality. (In *EDBED* they are called *Entropy* and *Capacity*; the names VC entropy and VC dimension as well as VC theory appeared later due to R. Dudley.)

[4]It is interesting to note that Fisher suggested along with the classical generative models (which he was able to justify), the heuristic solution (that belongs to a discriminative model) now called Fisher's linear discriminant function. This function minimizes some empirical loss functional, whose construction is similar to the Mahalanobis distance. For a long time this heuristic of Fisher was not considered an important result (it was ignored in most classical statistics textbooks). Only recently (after computers appeared and statistical learning theory became a subject not only of theoretical but also of practical justification) did Fisher's suggestion become a subject of interest.

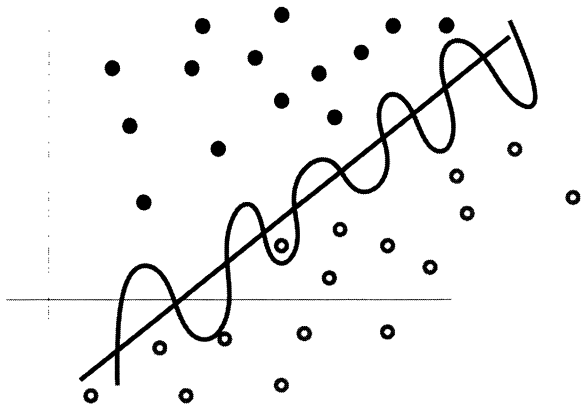

Figure 1.1: Two very different rules can make a similar classification.

Why do the generative and discriminative approaches lead to different results? There are two answers to this very important question which can be described from two different points of view: technical and philosophical (conceptual).

1.2.2 The Black Box Model

One can describe the pattern recognition problem as follows. There exists a black box BB that when given an input vector x_i returns an output y_i which can take only two values $y_i \in \{-1, +1\}$. The problem is: given the pairs $(y_i, x_i), i = 1, \ldots, \ell$ (the training data) find a function that approximates the rule that the black box uses.

Two different concepts of what is meant by a *good approximation* are possible:

(1) A good approximation of the BB rule is a function that is close (in a metric of functional space) to the function that the BB uses. (In the classical setting often we assume that the BB uses the Bayesian rule.)

(2) A good approximation of the BB rule is a function that provides approximately the same error rate as the one that the BB uses (provides the rule that predicts the outcomes of the BB well).

In other words, in the first case one uses a concept of closeness in the sense of being close to the *true function* used by the BB (closeness in a metric space of functions), while in the second case one uses a concept of closeness in the sense of being close to the accuracy of prediction (closeness in *functionals*). These definitions are very different.

In Figure 1.1 there are two different categories of data separated by two different rules. Suppose that the straight line is the function used by the black box. Then from the point of view of function estimation, the polynomial curve shown in Figure 1.1 is

very different from the line and therefore cannot be a good estimate of the *true BB* rule. From the other point of view, the polynomial rule separates the data well (and as we will show later can belong to a set with small VC dimension) and therefore can be a good *instrument* for prediction.

The lesson the Perceptron teaches us is that sometimes it is useful to give up the ambitious goal of estimating the rule the *BB* uses (the generative model of induction). Why?

Before discussing this question let me make the following remark. The problem of pattern recognition can be regarded as a generalization problem: using a set of data (observations) find a function[5] (theory). The same goals (but in more complicated situations) arise in the classical model of science: using observation of nature find the law. One can consider the pattern recognition problem as the simplest model of generalization where observations are just a set of i.i.d. vectors and the admissible laws are just a set of indicator functions. Therefore it is very useful to apply the ideas described in the general philosophy of induction to its simplest model and vice versa, to understand the ideas that appear in our particular model in the general terms of the classical philosophy. Later we will see that these interpretations are nontrivial.

1.2.3 Realism and Instrumentalism in the Philosophy of Science

The philosophy of science has two different points of view on the goals and the results of scientific activities.

(1) There is a group of philosophers who believe that the results of scientific discovery are the real laws that exist in nature. These philosophers are called the *realists*.

(2) There is another group of philosophers who believe the laws that are discovered by scientists are just an instrument to make a good prediction. The discovered laws can be very different from the ones that exist in Nature. These philosophers are called the *instrumentalists*.

The two types of approximations defined by classical discriminant analysis (using the generative model of data) and by statistical learning theory (using the function that explains the data best) reflect the positions of realists and instrumentalists in our simple model of the philosophy of generalization, the pattern recognition model. Later we will see that the position of philosophical instrumentalism played a crucial role in the success that pattern recognition technology has achieved.

However, to explain why this is so we must first discuss the theory of ill-posed problems, which in many respects describes the relationship between realism and instrumentalism in very clearly defined situations.

[5]The pattern recognition problem can be considered as the simplest generalization problem, since one has to find the function in a set of admissible *indicator* functions (that can take only two values, say 1 and -1.

1.3 Regularization and Structural Risk Minimization

1.3.1 Regularization of Ill-Posed Problems

In the beginning of the 1900s, Hadamard discovered a new mathematical phenomenon. He discovered that there are continuous operators A that map, in a one-to-one manner, elements of a space f to elements of a space F, but the inverse operator A^{-1} from the space F to the space f can be discontinuous. This means that there are operator equations

$$Af = F \tag{1.1}$$

whose solution in the set of functions $f \in \Phi$ exists, and is unique, but is unstable. (See Chapter 1 of EDED). That is, a small deviation $F + \Delta F$ of the (known) right-hand side of the equation can lead to a big deviation in the solution. Hadamard thought that this was just a mathematical phenomenon that could never appear in real-life problems. However, it was soon discovered that many important practical problems are described by such equations.

In particular, the problem of solving some types of linear operator equations (for example, Fredholm's integral equation of the second order) are ill-posed (see Chapter 1, Section 5 of *EDBED*). It was shown that many geophysical problems require solving (ill-posed) integral equations whose right-hand side is obtained from measurements (and therefore is not very accurate).

For us it is important that ill-posed problems can occur when one tries to estimate *unknown reasons from observed consequences*.

In 1943 an important step in understanding the structure of ill-posed problems was made. Tikhonov proved the so-called inverse operator lemma:

> *Let A be a continuous one-to-one operator from E_1 to E_2. Then the inverse operator A^{-1} defined on the images F of a compact set $f \in \Phi^*$ is stable.*

This means that if one possesses very strong prior knowledge about the solution (it belongs to a known compact set of functions), then it is possible to solve the equation. It took another 20 years before this lemma was transformed into specific approaches for solving ill-posed problems.

In 1962 Ivanov [21] suggested the following idea of solving operator equation (1.1). Consider the functional $\Omega(f) \geq 0$ that possesses the following two properties

(1) For any $c \geq 0$ the set of functions satisfying the constraint

$$\Omega(f) \leq c \tag{1.2}$$

is convex and compact.

(2) The solution f_0 of Equation (1.1) belongs to some compact set

$$\Omega(f_0) \leq c_0 \tag{1.3}$$

(where the constant $c_0 > 0$ may be unknown).

Under these conditions Ivanov proved that there exists a strategy for choosing $c = c(\varepsilon)$ depending on the accuracy of the right-hand side $||\Delta F||_{E_2} \leq \varepsilon$ such that the sequence of minima of the functional

$$R = ||Af - F||_{E_2} \tag{1.4}$$

subject to the constraints

$$\Omega(f) \leq c(\varepsilon) \tag{1.5}$$

converges to the solution of the ill-posed problem (1.1) as ε approaches zero.

In 1963 Tikhonov [55] proved the equivalent theorem that states: under conditions (1.2) and (1.3) defined on the functional $\Omega(f)$, there exists a function $\gamma_\varepsilon = \gamma(\varepsilon)$ such that the sequence of minima of the functionals

$$R_\gamma(f) = ||Af - F_\varepsilon||^2_{E_2} + \gamma_\varepsilon \Omega(f) \tag{1.6}$$

converges to the solution of the operator equation (1.1) as ε approaches zero.[6]

Both these results can be regarded as "*comforting ones*" since for any ε (even very small) one can guarantee nothing (the theorems guarantee only convergence of the sequence of solutions).

Therefore, one should try to avoid solving ill-posed problems by replacing them (if possible) with well-posed problems.

Keeping in mind the structure of ill-posed problems our problem of finding the *BB* solution can be split into two stages:

(1) Among a given set of admissible functions find a subset of functions that provides an expected loss that is close to the minimal one.

(2) Among functions that provide a small expected loss find one that is close to the *BB* function.

The first stage does not lead to an ill-posed problem, but the second stage might (if the corresponding operator is unstable).

The realist view requires solving both stages of the problem, while the instrumentalist view requires solving only the first stage and choosing for prediction any function that belongs to the set of functions obtained.

Technically, ill-posed problems appear in classical discriminant analysis as soon as one connects the construction of a discriminant function with the density estimation problem.

By definition, the density (if it exists) is a solution of the following equation

$$\int_a^x p(x')dx' = F(x), \tag{1.7}$$

[6]There is one more equivalent idea of how to solve ill-posed problems proposed in 1962 by Phillips [166]: minimize the functional $\Omega(f)$ satisfying the conditions defined above subject to the constraints

$$||Af - F||^2 \leq \varepsilon.$$

where $F(x)$ is a cumulative distribution function.

Therefore to estimate the density from the data

$$x_1, \ldots, x_\ell$$

means to solve Fredholm's equation (1.7) when the cumulative distribution function $F(x)$ is unknown but the data are given. One can construct an approximation to the unknown cumulative distribution function and use it as the right hand side of the equation. For example, one can construct the empirical distribution function

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i), \tag{1.8}$$

where

$$\theta(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{if } u < 0 \end{cases}.$$

It is known from Kolmogorov's bound for the Glivenco–Cantelli theorem that the empirical distribution function converges exponentially fast (not only asymptotically but for any set of fixed observations) to the desired cumulative distribution function. Using the empirical distribution function constructed from the data, one can try to solve this equation.

Note that this setting of the density estimation problem cannot be avoided since it reflects the definition of the density. Therefore in both parametric or nonparametric statistics, one has to solve this equation. The only difference is how the set of functions in which one is looking for the solution is defined: in a "narrow set of parametric functions" or in a "wide set of non-parametric functions".[7]

However, this point of view was not clearly developed in the framework of classical statistics, since both theories (parametric and nonparametric) of density estimation were constructed *before* the theory of solving ill-posed problems was introduced.

The general setting of the density estimation problem was described for the first time in *EDBED*. Later in Chapter 2, Section 2.3 when we discuss the SVM method, we will consider a pattern recognition problem, and show the difference between the solutions obtained by nonparametric statistics (based on the philosophy of realism) and by an SVM solution (based on the philosophy of instrumentalism).

Regularization Techniques

The regularization theory as introduced by Tikhonov suggests minimizing the equation

$$R_\gamma(f) = ||Af - F_\varepsilon||^2_{E_2} + \gamma_\varepsilon \Omega(f). \tag{1.9}$$

Under very specific requirements on the set of functions defined both by the functional $\Omega(f)$ and the value $c > 0$

$$\Omega(f) \leq c \tag{1.10}$$

[7]The maximum likelihood method suggested by Fisher is valid just for a very narrow admissible set of functions. It is already invalid, for example, for the set of densities defined by the sum of two Gaussians with unknown parameters (see example [139], Section 1.7.4.)

(for any $c > 0$ the set should be *convex and compact*), and under the condition that the desired solution *belongs to the set with some fixed* c_0, it is possible to define a strategy of choosing the values of the parameter γ that asymptotically lead to the solution.

1.3.2 Structural Risk Minimization

The Structural Risk Minimization (SRM) principle generalizes the Ivanov scheme in two ways:

(1) It considers a structure on any sets of functions (not necessarily defined by inequality (1.5)).

(2) It does not require compactness or convexity on the set of functions that define the element of the structure. It also does not require the desired solution belonging to one of the elements of the structure.

The only requirement is that every element of the nested sets possesses a finite VC dimension (or other capacity factor).

Under these general conditions the risks provided by functions that minimize the VC bound converge to the smallest possible risk (even if the desired function belongs to the closure of the elements). Also, for any fixed number of observations it defines the smallest guaranteed risk.

In the early 1970s Chervonenkis and I introduced SRM for sets of indicator functions (used in solving pattern recognition problems) [13]. In *EDBED* the SRM principle was generalized for sets of real-valued functions (used in solving regression estimation problems).

Therefore the difference between regularization and structural risk minimization can be described as follows.

> Regularization was introduced for solving ill-posed problems. It requires strong knowledge about the problem to be solved (the solution has to belong to the compact (1.10) defined by some constant c) and (generally speaking) does not have guaranteed bounds for a finite number of observations.
>
> Structural risk minimization was introduced for solving predictive problems. It is more general (does not require strong restrictions of admissible set of functions) and has a guaranteed bound for a finite number of observations.

Therefore if the regularization method is the main instrument for solving ill-posed problems using the *philosophical realism* approach, then the structural risk minimization method is the main instrument for solving problems using the *philosophical instrumentalism* approach.

Remark. In the late 1990s the concept of regularization started to be used in the general framework of minimizing the functionals (1.9) to solve predictive generalization problems. The idea was that under any definition of the functional $\Omega(f)$ there exists a parameter γ which leads to convergence to the desired result. This is, however,

incorrect: first, it depends on the concept of convergence; second, there are functionals (for which the set of functions (1.5) can violate finiteness of capacity conditions) that do not lead to convergence in any sense.

1.4 The Beginning of the Split Between Classical Statistics and Statistical Learning Theory

The philosophy described above was more or less clear by the end of the 1960s.[8] By that time there was no doubt that in analyzing the pattern recognition problem we came up with a new direction in the theory of generalization. The only question that remained was how to describe this new direction. Is this a new branch of science or is it a further development in classical statistics? This question was the subject of discussions in the seminars at the Institute of Control Sciences of the Academy of Sciences of USSR (Moscow).

The formal decision, however, was made when it came time to publish these results in the *Reports of Academy of Sciences of USSR* [143]. The problem was in which section of *Reports* it should be published — in "Control Sciences (Cybernetics)" or in "Statistics". It was published as a contribution in the "Control Sciences" section.

This is how one of the leading statisticians of the time, Boris Gnedenco, explained why it should not be published in the "Statistics" section:

> It is true that this theory came from the same roots and uses the same formal tools as statistics. However, to belong to the statistical branch of science this is not enough. Much more important is to share the same belief in the models and to share the same philosophy. Whatever you are suggesting is not in the spirit of what I am doing or what A. Kolmogorov is doing. It is not what our students are doing nor will it be what the students of our students do. Therefore, you must have your own students, develop your own philosophy, and create your own community.

More than 35 years have passed since this conversation. The more time passed, the more impressed I became with Gnedenco's judgment. The next three decades (1970s, 1980s, and 1990s) were crucial for developments in statistics. After the shocking discovery that the classical approach suffers from the curse of dimensionality, statisticians tried to find methods that could replace classical methods in solving real-life problems. During this time statistics was split into two very different parts: theoretical statistics that continued to develop the classical paradigm of generative models, and applied statistics that suggested a compromise between theoretical justification of the algorithms and heuristic approaches to solving real-life problems. They tried to justify such a position by inventing special names for such activities (exploratory data analysis), where in fact the superiority of common sense over theoretical justification was declared. However, they never tried to construct or justify new algorithms using VC

[8] It was the content of my first book *Pattern Recognition Problem* published in 1971 (in Russian).

theory. Only after SVM technology became a dominant force in data mining methods did they start to use its technical ideas (but not its philosophy) to modify classical algorithms.[9]

Statistical learning theory found its home in computer science. In particular, one of the most advanced institutions where SLT was developing in the 1970s and 1980s was the Institute of Control Sciences of the Academy of Sciences of USSR. Three different groups, each with different points of view on the generalization problem, became involved in such research: the Aizerman–Braverman–Rozonoer's group, the Tsypkin group, and the Vapnik–Chervonenkis group.

Of these groups ours was the youngest: I just got my PhD (candidate of science) thesis, and Chervonenkis got his several years later. Even so, our research direction was considered one of the most promising. In order to create a VC community I was granted permission from the Academy of Sciences to have my own PhD students.[10]

From this beginning we developed a statistical learning community. I had several very strong students including Tamara Glaskov, Anatoli Mikhalsky, Anatoli Stehanuyk, Alexander Sterin, Fellix Aidu, Sergey Kulikov, Natalia Markovich, Ada Sorin, and Alla Juravel who developed both machine learning theory and effective machine learning algorithms applied to geology and medicine.

By the end of the 1960s my department head, Alexander Lerner, made an extremely important advance in the application of machine learning: he convinced the high-level bureaucrats to create a laboratory for the application of machine learning techniques in medicine.

In 1970 such a laboratory was created in the State Oncology Centre. The director of the laboratory was my former PhD student, Tamara Glaskov.

It is hard to overestimate how much this laboratory accomplished during this time. Only recently have the most advanced oncology hospitals in the West created groups to analyze clinical data. This was routine in USSR decades earlier.

In beginning of the 1970s I prepared my doctoral thesis.

1.5 The Story Behind This Book

Government control under the Soviet Communist regime was total. One of its main modus operandi was to control who was promoted into more or less prominent positions. From the government bureaucrat's perspective a scientific degree (and especially a doctoral degree) holder possessed influence, and therefore they wanted to control who obtained this degree.

The execution of such control was one of the obligations of the institution called

[9]Statisticians did not recognise conceptual aspects of VC theory. Their criticism of this theory before SVM was that the VC bounds were too loose to be useful. Therefore the theory is not practical and to create new methods it is better to use common sense than the results of this theory.

[10]In the Russian system there were two academic degrees: *candidate of science* (which is equivalent to the PhD degree in the United States) and *doctor of science* (which is equivalent to the *Habilitation a Diriger des Recherches (HDR)* in France). Normally only doctors of science could have PhD students. I was granted this privilege and had to defend my doctoral thesis soon.

the Supreme Certifying Commission[11] (SCC) closely related to the KGB. The rule was that any decision on any thesis defense made by any Scientific Councils anywhere in the country must be approved by this commission. If the SCC disapproved several decisions by a particular Scientific Council it could be dismissed. Therefore the normal policy of academic institutions was not to enter into conflict with the SCC.

From the KGB's point of view I was a wrong person to obtain the doctoral level: I was not a member of the Communist Party, I was Jewish, my PhD adviser, Alexander Lerner, had applied for immigration to Israel and became a "refusenik," some of my friends were dissidents, and so on.

In this situation everybody understood that the Institute would be in conflict with the SCC's mandate. Nevertheless the feeling was that the support of the scientific community would be so strong that the SCC would not start the battle.

The SCC, however, reacted with a trick that to my knowledge was never used before: it requested that the Scientific Council change one of the reviewers to their trusted man who did his job: wrote a negative review.

I had a long conversation with the Chairman of the Scientific Council, Yakov Tsypkin, after he discussed the situation with the members of the Council. He told me that everyone on the Scientific Council understood what was going on and if I decided to defend my thesis the Scientific Council would unanimously support me. However, I had no chance of being approved by the SCC since they would have a formal reason to reject my thesis. Also they would have a formal reason to express distrust of the Scientific Council of the Institute. In this situation the best solution was to withdraw my thesis and publish it as a book. However, since the names of the authors of books were also under the KGB's control (the authors should also be "good guys") I would only be able to publish the book if my name did not attract too much attention. This would allow the editor, Vladimir Levantovsky (who was familiar with this story), to successfully carry out all necessary procedures to obtain permission (from the institution that controls the press) to publish the book.

So, I withdrew my thesis, rewrote it as a book, and due to the strong support of many scientists (especially Tsypkin), the editor Levantovsky was able to publish it (in Russian) in 1979.

In 1982 the well known American statistician, S. Kotz, translated it into English under the title *Estimation of Dependencies Based on Empirical Data* which was published by Springer. The first part of this volume is its reprint.

The main message that I tried to deliver in the book was that classical statistics could not overcome the curse of dimensionality but the new approach could. I devoted three chapters of the book to different classical approaches and demonstrated that none of them could overcome the curse of dimensionality. Only after that did I describe the new theory.

[11] The Russian abbreviation is VAK.

Chapter 2

FALSIFIABILITY AND PARSIMONY: VC DIMENSION AND THE NUMBER OF ENTITIES (1980–2000)

2.1 SIMPLIFICATION OF VC THEORY

For about ten years this book did not attract much attention either in Russia or in the West. It attracted attention later.

In the meantime, in 1984 (five years after the publication of the original version of this book and two years after its English translation) an important event happened. Leslie Valiant published a paper where he described his vision of how learning theory should be built [122].

Valiant proposed the model that later was called the *Probably Approximately Correct* (PAC) learning model. In this model, the goal of learning is to find a rule that reasonably well approximates the best possible rule. One has to construct algorithms which guarantee that such a rule will be found with some probability (not necessarily one). In fact, the PAC model is one of the major statistical models of convergence, called consistency. It has been widely used in statistics since at least Fisher's time.

Nevertheless Valiant's article was a big success. In the mid-1980s the general machine learning community was not very well connected to statistics. Valiant introduced to this community the concept of consistency and demonstrated its usefulness. The theory of consistency of learning processes as well as generalization bounds was the subject of our 1968 and 1971 articles [143, 11], and was described in detail in our 1974 book [12, 173] devoted to pattern recognition, and in a more general setting in *EDBED*. However, at that time these results were not well known in the West.[1]

[1] In 1989 I met Valiant in Santa Cruz, and he told me that he did not know of our results when he wrote

In the 20th century, and especially in the second half of it, mass culture began to play an important role. For us it is important to discuss the "scientific component" of mass culture.

With the increasing role of science in everyday life, the general public began to discuss scientific discoveries in different areas: physical science, computer science (cybernetics), cognitive science (pattern recognition), biology (genomics), and philosophy. The discussions were held using very simplified scientific models that could be understood by the masses. Also scientists tried to appeal to the general public by promoting their philosophy using simplified models (for example, as has been done by Wiener). There is nothing wrong with this.

However, when science becomes a mass profession, the elements of the scientific mass culture in some cases start to substitute for the real scientific culture: It is much easier to learn the slogans of the scientific mass culture than it is to learn many different concepts from the original scientific sources. Science and "scientific mass culture," however, are built on very different principles. In *Mathematical Discoveries*, Polya describes the principle of creating scientific mass culture observed by the remarkable mathematician Zermello. Here is the principle:

Gloss over the essentials and attract attention to the obvious.

Something that could remind this principle happened when (after appearance Valiant's article) the adaptation of ideas described in EDBED started. In the PAC adaptation the VC theory was significantly simplified by removing its essential parts.

In *EDBED* the main idea was the necessary and sufficient aspects of the theory based on three capacity concepts: the VC entropy, the Growth function, and the VC dimension. It stresses that the most accurate bounds can be obtained based on the VC entropy concept. This, however, requires information about the probability measure. One can construct less accurate bounds that are valid for all probability measures. To do this one has to calculate the Growth function which can have a different form for different sets of admissible functions. The Growth function can be upper bounded by the standard function that depends on only one integer parameter (the VC dimension). This also decreases accuracy, but makes the bounds simpler.

These three levels of the theory provide different possibilities for further developments in learning technology. For example, one can try to create theory for the case when the probability measure belongs to some specific sets of measures (say smooth ones), or one can try to find a better upper bound for the Growth function using a standard function that depends on say two (or more) parameters. This can lead to more accurate estimates and therefore to more advanced algorithms. The important component of the theory described in *EDBED* was the structural risk minimization principle. It was considered to be the main driving force behind predictive learning technology.

PAC theory started just from the definition of the VC dimension based on the combinatorial lemma used to estimate the bound for the Growth function (see *EDBED*, Chapter 6, Section A2). The main effort was placed on obtaining VC type bounds for

his article, and that he even visited a conference at Moscow University to explain this to me. Unfortunately we never met in Moscow. After his article was published Valiant tried to find the computer science aspects of machine learning research suggesting analyzing the computational complexity of learning problems. In 1990 he wrote [123]: "If the computational requirements is removed from the definition then we are left with the notion of non-parametric inference in sense of statistics as discussed in particular by Vapnik [*EDBED*]."

different classes of functions (say for neural networks), and on the generalizations of the theory for the set of nonindicator functions. In most cases these generalizations were based on extensions of the VC dimension concept for real-valued functions made in the style described in *EDBED*. The exception was the fat-shattering concept [141] related to VC entropy for real-valued functions described in Chapter 7.

In the early 1990s, some PAC researchers started to attack the VC theory. First, the VC theory was declared a "worst-case theory" since it is based on the uniform convergence concept. In contrast to this "worst-case-theory" the development of "real-case theory" was announced. However, this is impossible (see Section 1.2 of this Afterword) since the (one-sided) uniform convergence forms the necessary and sufficient conditions for consistency of learning (that is also true for PAC learning). Then in the mid-1990s an attempt was made to rename the Vapnik–Chervonenkis lemma (*EDBED*, Chapter 6, Sections 8 and A2) as the Sauer lemma. For the first time we published the formulation of this lemma in 1968 in the *Reports of the Academy of Sciences of USSR* [143]. In 1971, we published the corresponding proofs in the article devoted to the uniform law of large numbers [11]. In 1972, two mathematicians N. Sauer [130] and S. Shelah [131] independently proved this combinatorial lemma.

Researchers, who in the 1980s learned from *EDBED* (or from our articles) both the lemma and its role in statistical learning theory, renamed it in the 1990s.[2] Why?

My speculation is that renaming it was important for the dilution of VC theory and creating the following legend:

> In 1984 the PAC model was introduced. Early in statistics a concept called the VC dimension was developed. This concept plays an important role in the Sauer lemma, which is a key instrument in PAC theory.

Now, due to new developments in the VC theory and the interest in the advanced topics of statistical learning theory, this legend has died, and as a result interest in PAC theory has significantly decreased. This is, however, a shame because the computational complexity aspects of learning stressed by Valiant remain relevant.

2.2 Capacity Control

2.2.1 Bell Labs

In 1990 Larry Jackel, the head of the Adaptive Systems Research Department at AT&T Bell Labs, invited me to spend half a year with his group. It was a time of wide discussions on the VC dimension concept and its relationship to generalization problems. The obvious interpretation of the VC dimension was the number of free parameters that led to the curse of dimensionality. John Denker, a member of this department, showed,

[2] N. Sauer did not have in mind statistics proving this lemma. This is the content of the abstract of his article: "P. Erdös (oral communication) transmitted to me in Nice the following question: . . . (*the formulation of the lemma*). . . .. In this paper we will answer this question in the affirmative by determining the exact upper bounds."

however, that the VC dimension is not necessarily the number of free parameters. He came up with the example

$$y = \theta\{\sin ax\},\ x \in R^1,\ a \in (0, \infty)$$

a set of indicator functions that has only one free parameter yet possesses an infinite VC dimension (see Section 2.7.5, footnote 7). In *EDBED* another situation was described: when the VC dimension was smaller than the number of free parameters. These intriguing facts could lead to new developments in learning theory.

Our department had twelve researchers. Six of them, L. Jackel, J. Denker, S. Solla, C. Burges, G. Nohl, and H.P. Graf were physicists, and six, Y. LeCun, L. Bottou, P. Simard, I. Guyon, B. Boser, and Y. Bengio were computer scientists. The main direction of research was to advance the understanding of pattern recognition phenomena. To do this they relied on the principles of research common in physics.

The main principle of research in physics can be thought of as the complete opposite of the Zermello principle for creating scientific mass culture. It can be formulated as follows:

Find the essential in the nonobvious.

The entire story of creating modern technology can be seen as an illustration of this principle. At the time when electricity, electromagnetic waves, annihilation, and other physical fundamentals were discovered they seemed to be insignificant elements of nature. It took a lot of joint efforts of theorists, experimental physicists and engineers to prove that these negligible artifacts are very important parts of nature and make it work.

The examples given by Denker and another one described in *EDBED* (see Chapter 10, Section 5) could be an indication that such a situation in machine learning is quite possible.

The goal of our department was to understand and advance new general principles of learning that are effective for solving real-life problems. As a model problem for ongoing experiments, the department focused on developing automatic systems that could read handwritten digits. This task was chosen for a number of reasons. First, it was known to be a difficult problem, with traditional machine vision approaches making only slow progress. Second, lots of data were available for training and testing. And third, accurate solutions to the problem would have significant commercial importance.

Initial success in our research department led to the creation of a development group supervised by Charlie Stenard. This group, which worked closely with us, had as a goal the construction of a machine for banks that could read handwritten checks from all over the world. Such a machine could not make too many errors (the number of errors should be comparable to the number made by humans). However, the machine could refuse to read some percentage of checks.

I spent ten years with this department. During this time check reading machines became an important instrument in the banking industry. About 10% of checks in US banks are read by technology developed at Bell Labs.

During these years the performance of digit recognition was significantly improved. However, it *never* happened that significant improvements in quality of classification were the results of smart engineering heuristics. All jumps in performance were results of advances in understanding fundamentals of the pattern recognition problem.

2.2.2 Neural Networks

When I joined the department, the main instrument for pattern recognition was neural networks constructed by Yann LeCun, one of the originators of neural networks. For the digit recognition problem he designed a series of convolutional networks called LeNet. In the early 1990s this was a revolutionary idea. The traditional scheme of applying pattern recognition techniques was the following: a researcher constructs several very carefully crafted features and uses them as inputs for a statistical parametric model. To construct the desired rule they estimated the parameters of this model. Therefore good rules in many respects reflected how smart the researcher was in constructing features.

LeNet uses as input a high-dimensional vector whose coordinates are the raw image pixels. This vector is processed using a multilayer convolutional network with many free parameters. Using the back propagation technique, LeNet tunes the parameters to minimize the training loss.[3]

For the digit recognition problem, the rules obtained by LeNet were significantly better than any rules obtained by the classical style algorithms. This taught a great lesson: one does not need to go into the details of the decision rule; it is enough to create an "appropriate architecture" and an "appropriate minimization method" to solve the problem.

2.2.3 Neural Networks: The Challenge

The success of neural nets in solving pattern recognition problems was a challenge for theorists. Here is why. When one is trying to understand how the brain is working two different questions arise:

(1) What happens? What are the principles of generalization that the brain executes?

(2) How does it happen? How does the brain execute these principles?

Neural networks attempt to answer the second question using an artificial brain model motivated by neurophysiologists.

According to the VC theory, however, this is not very important. VC theory declares that two and only two factors are responsible for generalization. They are the value of empirical loss, and the capacity of the admissible set of functions (the VC entropy, Growth function, or the VC dimension). The SRM principle states that any method that controls these two factors well (minimizing the right-hand side of the VC bounds) is strongly universally consistent.

It was clear that artificial neural networks executed the structural risk minimization principle. However, they seemed to do this rather inefficiently. Indeed, the loss function that artificial neural networks minimize has many local minima. One can guarantee convergence to one of these minima but cannot guarantee good generalization. Neural networks practitioners define some initial conditions that they believe will lead to a

[3] As computer power increased, LeCun constructed more powerful generations of LeNet.

"good" minimum. Also, the back-propagation method based on the gradient procedure of minimization in high-dimensional spaces requires a very subtle treatment of step values. The choice of these values does not have a good recommendation.

In order to control capacity the designer chooses an appropriate number of elements (neurons) for the networks. Therefore for different training data sizes one has to design different neural networks. All these factors make neural networks more of an art than a science.

Several ideas that tried to overcome the described shortcomings of neural networks were checked during 1991 and 1992 including measuring the VC dimension (capacity) of the learning machine [144, 142] and constructing local learning rules [145]. Now these ideas are developing in a new situation. However, in 1992 they were overshadowed by a new learning concept called Support Vector Machines (SVMs).

2.3 Support Vector Machines (SVMs)

The development of SVMs has a 30-year history, from 1965 until 1995. It was completed in three major steps.

2.3.1 Step One: The Optimal Separating Hyperplane

In 1964, Chervonenkis and I came up with an algorithm for constructing an optimal separating hyperplane called the generalized portrait method. Three chapters of our 1974 book *Theory of pattern recognition*, contain the detailed theory of this algorithm [12, 173]. In *EDBED* (Addendum I), a simplified version of this algorithm is given. Here are more details. The problem was: given the training data

$$(y_1, x_1), \ldots, (y_\ell, x_\ell), \tag{2.1}$$

construct the hyperplane

$$(w_0, x) + b_0 = 0 \tag{2.2}$$

that separates these data and has the largest margin. In our 1974 book and in *EDBED* we assumed that the data were separable. The generalization of this algorithm for constructing an optimal hyperplane in the nonseparable case was introduced in 1995 [132]. We will discuss it in a later section.

Thus, the goal was to maximize the functional

$$\rho_0 = \inf_{\{i:\, y_i=1\}} \left[\left(\frac{w}{|w|}, x_i\right) + b\right] - \sup_{\{j:\, y_j=-1\}} \left[\left(\frac{w}{|w|}, x_j\right) + b\right]$$

under the constraints

$$y_i((w, x_i) + b) \geq 1, \quad i = 1, \ldots, \ell. \tag{2.3}$$

It is easy to see that this problem is equivalent to finding the minimum of the quadratic form

$$R_1(w, b) = (w, w)$$

subject to the linear constraints (2.3). Let this minimum be achieved when $w = w_0$. Then

$$\rho_0 = \frac{2}{\sqrt{(w_0, w_0)}}.$$

To minimize the functional (w, w) subject to constraints (2.3) the standard Lagrange optimization technique was used. The Lagrangian

$$L(\alpha) = \frac{1}{2}(w, w) - \sum_{i=1}^{\ell} \alpha_i([y_i((w, x_i) + b) - 1]) \tag{2.4}$$

(where $\alpha_i \geq 0$ are the Lagrange multipliers) was constructed and its minimax (minimum over w and b and maximum over the multipliers $\alpha_i \geq 0$) was found. The solution of this quadratic optimization problem has the form

$$w_0 = \sum_{i=1}^{\ell} y_i \alpha_i^0 x_i. \tag{2.5}$$

To find these coefficients one has to maximize the functional:

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (x_i, x_j) \tag{2.6}$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \ \alpha_i \geq 0, \ \ i = 1, \ldots, \ell.$$

Substituting (2.5) back into (2.2) we obtain the separating hyperplane expressed in terms of the Lagrange multipliers

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 (x, x_i) + b_0 = 0. \tag{2.7}$$

2.3.2 The VC Dimension of the Set of ρ-Margin Separating Hyperplanes

The following fact plays an important role in SVM theory. Let the vectors $x \in R^n$ belong to the sphere of radius $R = 1$. Then the VC dimension h of the set of hyperplanes with margin $\rho_0 = (w_0, w_0)^{-1}$ has the bound

$$h \leq \text{in} \quad \{(w_0, w_0), n\} + 1.$$

That is, the VC dimension is defined by the smallest of the two values: the dimensionality n of the vectors x and the value (w_0, w_0). In Hilbert (infinite dimensional) space,

the VC dimension of the set of separating hyperplanes with the margin ρ_0 depends just on the value (w_0, w_0).

In *EDBED* I gave a geometrical proof of the bound (See Chapter 10, Section 5). In 1997, Gurvits found an algebraic proof [124]. Therefore, the optimal separating hyperplane executes the SRM principle: it minimizes (to zero) the empirical loss, using the separating hyperplane that belongs to the set with the smallest VC dimension.

One can therefore introduce the following learning machine that executes the SRM principle:

> Map input vectors $x \in X$ into (a rich) Hilbert space $z \in Z$, and construct the maximal margin hyperplane in this space.

According to the VC theory the generalization bounds depend on the VC dimension. Therefore by controlling the margin of the separating hyperplane one controls the generalization ability.

2.3.3 Step Two: Capacity Control in Hilbert Space

The formal implementation of this idea requires one to specify the operator

$$z = \mathcal{F}x$$

which should be used for mapping. Then similar to (2.7) one constructs the separating hyperplane in image space

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 (z, z_i) + b_0 = 0,$$

where the coefficients $\alpha_i \geq 0$ are the ones that maximize the quadratic form

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (z_i, z_j) \tag{2.8}$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \ \alpha_i \geq 0, \quad i = 1, \ldots, \ell. \tag{2.9}$$

In 1992 Boser, Guyon and I found an effective way to construct the optimal separating hyperplane in Hilbert space without explicitly mapping the input vectors x into vectors z of the Hilbert space [125].

This was done using Mercer's theorem.

> *Let vectors $x \in X$ be mapped into vectors $z \in Z$ of some Hilbert space.*
>
> **1.** *Then there exists in X space a symmetric positive definite function $K(x_i, x_j)$ that defines the corresponding inner product in Z space:*
>
> $$(z_i, z_j) = K(x_i, x_j).$$

2. *Also, for any symmetric positive definite function $K(x_i, x_j)$ in X space there exists a mapping from X to Z such that this function defines an inner product in Z space.*

Therefore, according to Mercer's theorem, the separating hyperplane in image space has the form

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 K(x, x_i) + b_0 = 0,$$

where the coefficients α_i^0 are defined as the solution of the quadratic optimization problem: maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.10}$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \ \alpha_i \geq 0, \quad i = 1, \ldots, \ell. \tag{2.11}$$

Choosing specific kernel functions $K(x_i, x_j)$ one makes specific mappings from input vectors x into image vectors z.

The idea of using Mercer's theorem to map into Hilbert space was used in the mid-1960s by Aizerman, Braverman, and Rozonoer [2]. Thirty years later we used this idea in a wider context.

2.3.4 Step Three: Support Vector Machines

In 1995 Cortes and I generalized the maximal margin idea for constructing (in image space) the hyperplane

$$(w_0, z) + b_0 = 0$$

when the training data are nonseparable [132]. This technology became known as Support Vector Machines (SVMs). To construct such a hyperplane we follow the recommendations of the SRM principle.

Problem 1. Choose among the set hyperplanes with the predefined margin

$$\rho^2 = \frac{4}{(w_0, w_0)} \leq H = \frac{1}{h}$$

the one that separates the images of the training data with the smallest number of errors. That is, we minimize the functional

$$R = \sum_{i=1}^{\ell} \theta(\xi_i) \tag{2.12}$$

subject to the constraints

$$y_i((w, z_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, \ell \tag{2.13}$$

and the constraint

$$(w, w) \leq h, \tag{2.14}$$

where $\theta(u)$ is the step function:

$$\theta(u) = \begin{cases} 1, & \text{if } u \geq 0 \\ 0, & \text{if } u < 0. \end{cases}$$

For computational reasons, however, we approximate *Problem 1* with the following one.

Problem 2. Minimize the functional

$$R = \sum_{i=1}^{\ell} \xi_i \tag{2.15}$$

(instead of the functional (2.12)) subject to the constraints (2.13) and (2.14).

Using the Lagrange multiplier technique, one can show that the corresponding hyperplane has an expansion

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 (z_i, z) + b_0 = 0. \tag{2.16}$$

To find the multipliers one has to maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - h\sqrt{\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (z_i, z_j)} \tag{2.17}$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{2.18}$$

and the constraints

$$0 \leq \alpha_i \leq 1, \quad i = 1, \ldots, \ell.$$

Problem 3. Problem 2 is equivalent to the following (reparametrized) one: Minimize the functional

$$R = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} \xi_i \tag{2.19}$$

subject to constraints (2.13). This setting implies the following dual space solution: Maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (z_i, z_j) \tag{2.20}$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{2.21}$$

and the constraints

$$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, \ell.$$

One can show that for any h there exists a C such that the solutions of Problem 2 and Problem 3 coincide. From a computational point of view Problem 3 is simpler than Problem 2. However, in Problem 2 the parameter h estimates the VC dimension. Since the VC bound depends on the ratio h/ℓ one can choose the VC dimension to be some fraction of the training data, while in the reparametrized Problem 3 the corresponding parameter C cannot be specified; it can be any value depending on the VC dimension and the particular data.

Taking into account Mercer's theorem,

$$(z_i, z_j) = K(x_i, x_j),$$

we can rewrite the nonlinear separating rule in input space X as

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x) + b_0 = 0, \tag{2.22}$$

where the coefficients are the solution of the following problems:

Problem 1a. Minimize the functional

$$R = \sum_{i=1}^{\ell} \theta(\xi_i) \tag{2.23}$$

subject to the constraints

$$y_i \sum_{j=1}^{\ell} (y_j \alpha_j K(x_j, x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, \ell \tag{2.24}$$

and the constraint

$$\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \leq h. \tag{2.25}$$

Problem 2a. Maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - h \sqrt{\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j)} \tag{2.26}$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{2.27}$$

and the constraints

$$0 \le \alpha_i \le 1, \quad i = 1, \ldots, \ell. \tag{2.28}$$

Problem 3a. Maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \tag{2.29}$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0$$

and the constraints

$$0 \le \alpha_i \le C, \quad i = 1, ..., \ell.$$

The solution of Problem 3a became the standard SVM method. In this solution only some of the coefficients α_i^0 are different from zero. The vectors x_i for which $\alpha_i^0 \neq 0$ in (2.22) are called the *support vectors*. Therefore, the separating rule (2.22) is the expansion on the support vectors.

To construct a support vector machine one can use any (conditionally) positive definite function $K(x_i, x_j)$ creating different types of SVMs. One can even use kernels in the situation when input vectors belong to nonvectorial spaces. For example, the inputs may be sequences of symbols of different size (as in problems of bioinformatics or text classification). Therefore SVMs form a universal generalization engine that can be used for different problems of interest.

Two examples of Mercer kernels are the polynomial kernel of degree d

$$K(x_i, x_j) = ((x_i, x_j) + c)^d, \; c \ge 0 \tag{2.30}$$

and the exponential kernel

$$K(x_i, x_j) = \exp\left\{ - \left(\frac{|x_i - x_j|}{\sigma} \right)^d \right\}, \; \sigma > 0, \; 0 \le d \le 2. \tag{2.31}$$

2.3.5 SVMs and Nonparametric Statistical Methods

SVMs execute the idea of the structural risk minimization principle, where the choice of the appropriate element of the structure is defined by the constant C (and a kernel parameters). Therefore, theoretically, for any appropriate kernel (say for (2.31) by controlling parameters (which depends on the training data) one guarantees asymptotic convergence of the SVM solutions to the best possible solution [167].

In 1980 Devroye and Wagner proved that classical nonparametric methods of density estimation are also universally consistent [134]. That is, by controlling the parameter $\sigma_\ell = \sigma(\ell) > 0$ depending on the size ℓ of the training data, the following approximation of the density function

$$\bar{p}(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{1}{(2\pi)^{n/2}\sigma_\ell^n} \exp\left\{ -\left(\frac{|x_i - x|}{\sigma_\ell} \right)^2 \right\} \tag{2.32}$$

converges (in the uniform metric) to the desired density *with increasing* ℓ.

However, by choosing an appropriate parameter C of SVM, one controls the VC bound for any finite number of observations. One can also control these bounds by choosing the parameters of the kernels.

This section illustrates the practical advantage of this fact.

Let us use the nonparametric density estimation method to approximate the optimal (generative) decision rule for binary classification

$$p_1(x) - p_2(x) = 0, \tag{2.33}$$

where $p_1(x)$ is the density function of the vectors belonging to the first class and $p_2(x)$ is the density function of the vectors belonging to the second class. Here for notational simplicity we assume that the two classes are equally likely and that the number of training samples from the first and second class is the same. Using (2.32) the approximation (2.33) can be rewritten as follows.

$$\sum_{\{i:\, y_i=1\}} \exp\left\{ -\left(\frac{|x - x_i|}{\sigma} \right)^2 \right\} - \sum_{\{j:\, y_j=-1\}} \exp\left\{ -\left(\frac{|x - x_j|}{\sigma} \right)^2 \right\} = 0.$$

The SVM solution using the same kernel has the form

$$\sum_{\{i:\, y_i=1\}} \alpha_i \exp\left\{ -\left(\frac{|x - x_i|}{\sigma} \right)^2 \right\} - \sum_{\{j:\, y_j=-1\}} \alpha_j \exp\left\{ -\left(\frac{|x - x_j|}{\sigma} \right)^2 \right\} = 0.$$

Since our kernel is a positive definite function there exists a space Z where it defines an inner product (by the second part of Mercer's theorem). In Z space both solutions define separating hyperplanes

$$\sum_{\{i:\, y_i=1\}} (z_i, z) - \sum_{\{j:\, y_j=-1\}} (z_j, z) = 0$$

(the classical non-parametric solution) [152] and

$$\sum_{\{i:\, y_i=1\}} \alpha_i (z_i, z) - \sum_{\{j:\, y_j=-1\}} \alpha_j (z_j, z) = 0.$$

(the SVM solution). Figure 2.1 shows these solutions in Z space. The separating hyperplane obtained by nonparametric statistics is defined by the hyperplane orthogonal

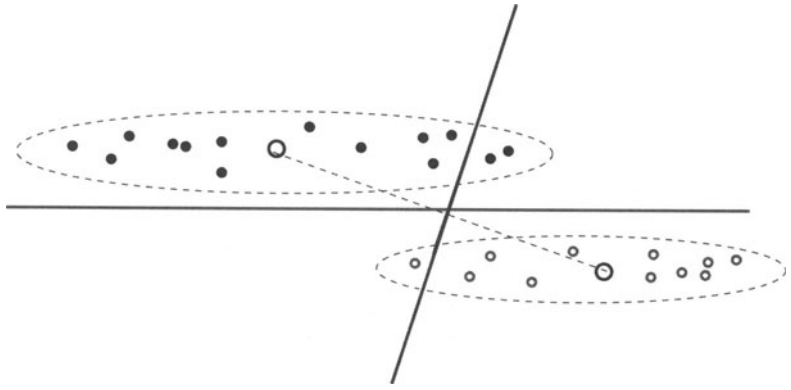

Figure 2.1: Classifications given by the classical nonparametric method and the SVM are very different.

to the line connecting the center of mass of two different classes. The SVM produces the optimal separating hyperplane.

In spite of the fact that both solutions converge asymptotically to the best one[4] they are very different for a fixed number of training data since the SVM solution is optimal (for any number of observations it guarantees the smallest predictive loss), while the non-parametric technique is not.

This makes SVM a state-of-the-art technology in solving real-life problems.

2.4 An Extension of SVMs: SVM+

In this section we consider a new algorithm called SVM+, which is an extension of SVM. SVM+ takes into account a known structure of the given data.

2.4.1 Basic Extension of SVMs

Suppose that our data are the union of $t \geq 1$ groups:

$$(X, Y)_r = (x_{r_1}, y_{r_1}), \ldots, (x_{r_{n_r}}, y_{r_{n_r}}), \quad r = 1, \ldots, t.$$

Let us denote indices from the group r by

$$T_r = \{i_{n_1}, \ldots, i_{n_r}\}, \quad r = 1, \ldots, t.$$

[4]Note that nonparametric density estimate (2.32) requires dependence of σ from ℓ. Therefore, it uses different Z spaces for different ℓ.

Let inside one group the slacks be defined by some correcting function that belongs to a given set of functions

$$\xi_i = \xi_r(x_i) = \phi_r(x_i, w_r), \quad w_r \in W_r, \ i \in T_r, \ r = 1, \ldots, t. \tag{2.34}$$

The goal is to define the decision function for a situation when sets of admissible correcting functions are restricted (when sets of admissible correcting functions are not restricted we are back to conventional SVM). By introducing groups of data and different sets of correcting functions for different groups one introduces additional information about the problem to be solved.

To define the correcting function $\xi(x) = \phi_r(x, w_r)$ for group T_r we map the input vectors $x_i, \ i \in T_r$ simultaneously into two different Hilbert spaces: into the space $z_i \in Z$ which defines the decision function (as we did for the conventional SVM) and into correcting function space $z_i^r \in Z_r$ which defines the set of correcting functions for a given group r. (Note that vectors of different groups are mapped into the same decision space Z but different correcting spaces Z_r.)

Let the inner products in the corresponding spaces be defined by the kernels

$$(z_i, z_j) = K(x_i, x_j), \ \forall i, j$$

and

$$(z_i^r, z_j^r) = K_r(x_i, x_j), \ i, j \in T_r, \ r = 1, \ldots, t. \tag{2.35}$$

Let the set of admissible correcting functions $\xi_r(x) = \phi_r(x, w_r), \ w_r \in W_r$, be linear in each Z_r space

$$\xi(x_i) = \phi_r(x, w_r) = [(w_r, z_i^r) + d_r] \geq 0, \ i \in T_r, \ r = 1, \ldots, t. \tag{2.36}$$

As before our goal is to find the separating hyperplane in decision space Z,

$$(w_0, z) + b_0 = 0$$

whose parameters w_0 and b_0 minimize the functional

$$R(w, w_1, \ldots, w_t) = \frac{1}{2}(w, w) + C \sum_{r=1}^{t} \sum_{i \in T_r} ((w_r, z_i^r) + d_r), \tag{2.37}$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - ((z_i^r, w_r) + d_r), \ i \in T_r, \ r = 1, \ldots, t \tag{2.38}$$

and the constraints

$$(w_r, z_i^r) + d_r \geq 0, \ i \in T_r, \ r = 1, \ldots, t. \tag{2.39}$$

Note that for set (2.36) the solution of this optimization problem does exist.

The corresponding Lagrangian is

$$L(w, w_1, \ldots, w_t; \alpha, \mu) = \frac{1}{2}(w, w) + C \sum_{r=1}^{t} \sum_{i \in T_r} ((w_r, z_i^r) + d_r) \tag{2.40}$$

$$-\sum_{i=1}^{\ell}\alpha_i[y_i((w,z_i)+b)-1+d_r+(w_r,z_i^r)]-\sum_{i=1}^{\ell}\mu_i((w_r,z_i^r)+d_r).$$

Using the same dual optimization technique as above one can show that the optimal separating hyperplane in Z space has the form

$$\sum_{i=1}^{\ell}\alpha_i^0 y_i(z_i,z)+b_0=0,$$

where the coefficients $\alpha_i^0 \geq 0$ minimize the same quadratic form as before

$$W(\alpha)=\sum_{i=1}^{\ell}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{\ell}y_iy_j\alpha_i\alpha_jK(x_i,x_j) \tag{2.41}$$

subject to the conventional constraint

$$\sum_{i=1}^{\ell}y_i\alpha_i=0 \tag{2.42}$$

and the new constraints

$$\sum_{i\in T_r}(\alpha_i+\mu_i)=|T_r|C,\ r=1,\ldots,t \tag{2.43}$$

($|T_r|$ is the number of elements in T_r),

$$\sum_{i\in T_r}(\alpha_i+\mu_i)K_r(x_i,x_j)=C\sum_{i\in T_r}K_r(x_i,x_j),\ j\in T_r,\ r=1,\ldots,t. \tag{2.44}$$

and constraints

$$\alpha_i\geq 0,\ \mu_i\geq 0,\quad i=1,\ldots,\ell.$$

When either

(1) There is no structure in the data: any vector belongs to its own group,

or

(2) There is no correlation between slacks inside all groups: $K_r(x_i,x_j)$ is an identity matrix for all r

$$K_r(x_i,x_j)=\begin{cases}1, & \text{if } i=j\\ 0, & \text{if } i\neq j\end{cases} \tag{2.45}$$

then Equation (2.44) defines the box constraints as in conventional SVMs (in case (2) Equations (2.43) are satisfied automatically). Therefore the SVM+ model contains the classical SVM model as a particular case.

The advantage of the SVM+ is the ability to consider the global structure of the problem that the conventional SVM ignores (see Section 2.4.3 for details).

This, however, requires solving a more general quadratic optimization problem to minimize in the space of 2ℓ nonnegative variables the same objective function subject to $(\ell+t+1)$ linear constraints (instead of one minimizing this objective function in the space of ℓ variable subjects of one linear constraint and ℓ box constraints in the conventional SVM).

2.4.2 Another Extension of SVM: SVM$_\gamma$+

Consider another extension of SVM, the so-called SVM$_\gamma$+, which directly controls the capacity of sets of correcting functions.

Let us instead of objective function (2.37) consider the function

$$R(w, w_1, \ldots, w_t) = \frac{1}{2}(w, w) + \frac{\gamma}{2}\sum_{r=1}^{t}(w_r, w_r) + C\sum_{r=1}^{t}\sum_{i\in T_r}((w_r, z_i^r) + d_r), \quad (2.46)$$

where $\gamma > 0$ is some value. When γ approaches zero (2.46) and (2.37) coincide.

The SVM$_\gamma$+ solution minimizes functional (2.46) subject to the constraints (2.38) and (2.39). To solve this problem we construct the Lagrangian. Comparing it to (2.40), this Lagrangian has one extra term $\gamma/2\sum(w_r, w_r)$. Repeating almost the same algebra as in the previous section we obtain that for the modified Lagrangian the dual space solution that defines the coefficients α_i^0 must maximize the functional

$$W(\alpha, \mu) = \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) +$$

$$\frac{C}{\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i + \mu_i)K_r(x_i, x_j) - \frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i + \mu_i)(\alpha_j + \mu_j)K_r(x_i, x_j)$$

subject to the constraints (2.42) and the constraints

$$\sum_{i\in T_r}(\alpha_i + \mu_i) = |T_r|C, \; r = 1, \ldots, t,$$

$$\alpha_i \geq 0, \; \mu_i \geq 0, \; i = 1, \ldots, \ell.$$

Note that when either:

(1) There is no structure (every training vector belongs to its own group),

or

(2) There is no correlation inside groups ((2.45) holds for all r) and $\gamma \longrightarrow 0$

then the SVM$_\gamma$+ solution coincides with the conventional SVM solution.

This solution requires maximizing the quadratic objective function in the space of 2ℓ nonnegative variables subject to $t + 1$ equality constraints.

One can simplify the computation when using models of correcting functions (2.36) with $d_r = 0, \; r = 1, ..., t$. In this case one has to maximize the functional $W(\alpha, \mu)$ over non-negative variables $\alpha_i \; \mu_i, \; i = 1, ..., \ell$ subject to one equality constraint (2.42).

2.4.3 Learning Hidden Information

SVM+ is an instrument for a new inference technology which can be called *Learning Hidden Information* (LHI). It allows one to extract additional information in situations where conventional technologies cannot be used.

WHAT INFORMATION CAN BE HIDDEN?

Consider the pattern recognition problem. Let one be given the training set

$$(x_1, y_1), \ldots, (x_\ell, y_\ell).$$

Suppose that one can add to this set additional information from two sources:

(1) information that exists in *hidden classifications* of the training set and

(2) information that exists in *hidden variables* of the training set.

The next two examples describe such situations.

EXAMPLE 1 (Information given in hidden classifications).
Suppose that one's goal is to find a rule that separates cancer patients from non cancer patients. One collects training data and assigns class $y_i = 1$, or $y_i = -1$, to patient x_i depending on the result of analysis of tissue taken during surgery. Analyzing the tissue, a doctor composes a report which not only concludes that the patient has a cancer (+1) or benign diagnosis (-1) but also that the patient belongs to a particular group (has a specific type of cancer or has a specific type of cell and so on). That is, the doctor's classification of the training data y_i^* is more detailed than the desired classification y_i. When constructing a classification rule $y = f(x)$, one can take into account information about y_i^*. This information can be used, for example, to create appropriate groups.

EXAMPLE 2 (Information given in hidden variables).
Suppose that one's goal is to construct a rule $y = f(x)$. However, for the training data along with the nonhidden variables x_i, one can determine the hidden variables x_i^*. The problem is using the data

$$(x_1, x_1^*, y_1), \ldots, (x_\ell, x_\ell^*, y_\ell)$$

which contain both nonhidden and hidden variables and their classifications y_i, to construct a rule $y = f(x)$ (rather than a rule $y = f(x, x^*)$) that makes a prediction based on nonhidden variables. By using variables x for a decision space and variables x, x^* for a correcting space one can solve this problem.

EXAMPLE 3 (Special rule for selected features).
A particular case of the problem described in Example 2 is constructing a decision rule for selected features, using information about the whole set of features. In this problem, the selected features are considered as non hidden variables while the rest of the features are hidden variables.

THE GENERAL PROBLEM
How should one construct (a more accurate than conventional) rule $y = f(x)$ using the data

$$(x_1, x_1^*, y_1, y_1^*), \ldots, (x_\ell, x_\ell^*, y_\ell, y_\ell^*)$$

instead of the data

$$(x_1, y_1), \ldots, (x_\ell, y_\ell).$$

To do this one can use the SVM+ method. Constructing the desired decision rule in the solution space, SVM+ uses two new ideas:

(1) It uses structure on training data and

(2) It uses several different spaces: (a) the solution space of nonhidden variables and (b) the correcting spaces of joint hidden and nonhidden variables.

SVM+ allows one to effectively use additional hidden information. The success of SVM+ depends on the quality of recovered hidden information.

The LHI technology using SVM+ requires the following three steps:

1. Use the data (x_i, x_i^*, y_i, y_i^*) for constructing a structure on the training set.

2. Use the kernel $K(x_i, x_j)$ for constructing a rule in the decision space, and

3. Use the kernels $K_r(x_i, x_i^* y_i^*; x_j, x_j^*, y_j^*)$ in the correcting spaces.

Note that in the SVM+ method the idea of creating a structure on the training set differs from the classical idea of clustering of the training set.

2.5 Generalization for Regression Estimation Problem

In this section we use the ε-insensitive loss function introduced in [140],

$$u_\varepsilon = \begin{cases} |u| - \varepsilon, & \text{if } |u| \geq \varepsilon \\ 0, & \text{if } |u| < \varepsilon. \end{cases}$$

This function allows one to transfer some properties of the SVM for pattern recognition (the accuracy and the sparsity) to the regression problem.

2.5.1 SVM Regression

Consider the regression problem: given iid data

$$(x_1, y_1), \ldots, (x_\ell, y_\ell),$$

where $x \in X$ is a vector and $y \in (-\infty, \infty)$ is a real value, estimate the function in a given set of real-valued functions.

As before using kernel techniques we map input vectors x into the space of image vectors $z \in Z$ and approximate the regression by a linear function

$$y = (w, z) + b, \tag{2.47}$$

where w and b have to be defined. Our goal is to minimize the following loss,

$$R = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} |y_i - (w, z) - b|_\varepsilon. \tag{2.48}$$

To minimize the functional (2.48) we solve the following equivalent problem [140]: Minimize the functional

$$R = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \tag{2.49}$$

subject to the constraints

$$y_i - (w, z_i) - b \leq \varepsilon + \xi_i^*, \quad \xi_i^* \geq 0, \quad i = 1, \ldots, \ell, \tag{2.50}$$

$$(w, z_i) + b - y_i \leq \varepsilon + \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, \ell. \tag{2.51}$$

To solve this problem one constructs the Lagrangian

$$L = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} \alpha_i [y_i - (w, z_i) - b + \varepsilon + \xi_i] \tag{2.52}$$

$$- \sum_{i=1}^{\ell} \alpha_i^* [(w, z_i) + b - y_i + \varepsilon + \xi_i^*] - \sum_{i=1}^{\ell} (\beta_i \xi_i + \beta_i^* \xi_i^*)$$

whose minimum over w, b, and ξ, ξ_i^* leads to the equations

$$w = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) z_i, \tag{2.53}$$

$$\sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0, \tag{2.54}$$

and

$$\alpha_i^* + \beta_i^* = C, \ \alpha_i + \beta_i = C, \tag{2.55}$$

where $\alpha, \ \alpha^*, \beta, \beta^* \geq 0$ are the Lagrange multipliers. Putting (2.53) into (2.47) we obtain that in X space the desired function has the kernel form

$$y = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) K(x_i, x) + b. \tag{2.56}$$

To find the Lagrange multipliers one has to put the obtained equation back into the Lagrangian and maximize the obtained expression.

Putting (2.53), (2.54), and (2.55) back into (2.52) we obtain

$$W = -\sum_{i=1}^{\ell} \varepsilon(\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i,j}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j). \tag{2.57}$$

To find $\alpha_i, \ \alpha_i^*$ for the approximation (2.56) one has to maximize this functional subject to the constraints

$$\sum_{i=1}^{\ell} \alpha_i^* = \sum_{i=1}^{\ell} \alpha_i,$$

$$0 \leq \alpha_i \leq C, \quad 0 \leq \alpha_i^* \leq C, \quad i = 1, \ldots, \ell.$$

2.5.2 SVM+ Regression

Now let us solve the same regression problem of minimizing the functional (2.49) subject to the constraints (2.50) and (2.51) in the situation when the slacks ξ_i and ξ_i^* are defined by functions from the set described in Section 2.4:

$$\xi_i = \phi_r(x_i, w_r) = (w_r, z_i) - d_r \geq 0,\ i \in T_r,\ \ r = 1, \ldots, t \tag{2.58}$$

$$\xi_i^* = \phi_r^*(x_i, w_r^*) = (w_r^*, z_i) - d_r^* \geq 0,\ i \in T_r,\ \ r = 1, \ldots, t. \tag{2.59}$$

To find the regression we construct the Lagrangian similar to (2.52) where instead of slacks ξ_i and ξ_i^* we use their expressions (2.58) and (2.59).

Minimizing this Lagrangian over w, b (as before) and over $w_r,\ d_r, w_r^*,\ d_r^*,\ r = 1, ..., t$ (instead of slacks ξ_i, and ξ_i^*) we obtain Equations (2.53) and (2.54) and the equations

$$\sum_{i \in T_r} (\alpha_i + \beta_i) z_i^r = C \sum_{i \in T_r} z_i^r,\ \sum_{i \in T_r} (\alpha_i^* + \beta_i^*) z_i^r = C \sum_{i \in T_r} z_i^r,\ r = 1, \ldots, t, \tag{2.60}$$

$$\sum_{i \in T_r} (\alpha_i^* + \beta_i^*) = C|T_r|,\ \sum_{i \in T_r} (\alpha_i + \beta_i) = C|T_r|,\ r = 1, \ldots, t \tag{2.61}$$

Putting these equations back into the Lagrangian we obtain

$$W = -\sum_{i=1}^{\ell} \varepsilon(\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i,j}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j). \tag{2.62}$$

From (2.60) and (2.61) we obtain

$$\sum_{i \in T_r} (\alpha_i + \beta_i) K_r(x_j, x_j) = C \sum_{i \in T_r} K_r(x_i, x_j),\ r = 1, \ldots, t,\ j \in T_r, \tag{2.63}$$

$$\sum_{i \in T_r} (\alpha_i^* + \beta_i^*) K_r(x_i, x_j) = C \sum_{i \in T_r} K_r(x_i, x_j),\quad r = 1, ..., t\ j \in T_r, \tag{2.64}$$

$$\alpha_i \geq 0,\ \ \beta_i \geq 0,\quad i = 1, \ldots, \ell.$$

Therefore to estimate the SVM+ regression function (2.56) one has to maximize the functional (2.62) subject to the constraints (2.54), (2.61), (2.63), (2.64).

2.5.3 SVM$_\gamma$+ Regression

Consider SVM$_\gamma$+ extension of regression estimation problem: Minimize the functional

$$R = \frac{1}{2}(w, w) + \frac{\gamma}{2} \left(\sum_{r=1}^{t} (w_r, w_r) + \sum_{r=1}^{t} (w_r^*, w_r^*) \right) + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \tag{2.65}$$

(instead of functional (2.49)) subject to constraints (2.50) and (2.51), where slacks ξ_i and ξ_i^* are defined by the correcting functions (2.58) and (2.59). The new objective function approaches (2.49) when γ approaches zero.

The same algebra of the Lagrange multiplier technique that was used above now implies that to find the coefficients α_i, α_i^* for approximation (2.56) one has to maximize the functional

$$W = -\sum_{i=1}^{\ell} \varepsilon(\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) +$$

$$\frac{C}{\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i + \beta_i)K_r(x_i, x_j) - \frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i + \beta_i)(\alpha_j + \beta_j)K_r(x_i, x_j) +$$

$$\frac{C}{\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i^* + \beta_i^*)K_r(x_i, x_j) - \frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i^* + \beta_i^*)(\alpha_j^* + \beta_j^*)K_r(x_i, x_j)$$

subject to the constraints

$$\sum_{i=1}^{\ell}\alpha_i^* = \sum_{i=1}^{\ell}\alpha_i,$$

$$\sum_{i\in T_r}(\alpha_i + \beta_i) = |T_r|C,\ r = 1, \ldots, t,$$

$$\sum_{i\in T_r}(\alpha_i^* + \beta_i^*) = |T_r|C,\ r = 1, \ldots, t,$$

$$\alpha_i \geq 0,\ \ \alpha_i^* \geq 0,\ \ \beta_i \geq 0,\ \ \beta_i^* \geq 0,\ \ \ i = 1, ..., \ell.$$

When either (1) there is no structure ($t = \ell$) or (2) there are no correlations ($K_r(x_i, x_j)$ has the form (2.45)) and $\gamma \to 0$ the solutions defined by SVM+ or SVM$_\gamma$+ regression coincide with the conventional SVM solution for regression.

2.6 The Third Generation

In the mid-1990s the third generation of statistical learning theory (SLT) researchers appeared. They were well-educated, strongly motivated, and hard working PhD students from Europe. Many European universities allow their PhD students to work on their theses anywhere in the world, and several such students joined our department in order to work on their thesis. First came Bernhard Schölkopf, Volker Blanz, and Alex Smola from Germany, then Jason Weston from England, followed by Olivier Chapelle, Olivier Bousquet, and Andre Eliseeff from France, Pascal Vincent from Canada, and Corina Cortes (PhD student from Rochester university). At that time support vector technology had just started to develop. Later many talented young people followed this direction but these were the first from the third generation of researchers.

I would like to add to this group two young AT&T researchers of that time: Yoav Freund and Robert Schapire, who did not directly follow the line of statistical learning theory and developed *boosting* technology that is close to the one discussed here [135, 136].

The third generation transformed both the area of machine learning research and the style of research. During a short period of time (less than ten years) they created a new direction in statistical learning theory: SVM and kernel methods. The format of this Afterword does not allow me to go into details of their work (there are hundreds of first-class articles devoted to this subject and it is very difficult to choose from them). I will just quote some of their textbooks [152–158], collective monographs and workshop materials [159–164]. Also I would like to mention the tutorial by Burges [165] which demonstrated the unity of theoretical and algorithmical parts of VC theory in a simple and convincing way.

The important achievement of the third generation was creating a large international SVM (kernel) community. They did it by accomplishing three things:

(1) Constructing and supporting a special Website called Kernel Machine (www.kernel-machines.org).

(2) Organizing eight machine learning workshops and five Summer Schools, where advanced topics relevant to empirical inference research were taught. These topics included:
— Statistical learning theory,
— Theory of empirical processes,
— Functional analysis,
— Theory of approximation,
— Optimization theory, and
— Machine learning algorithms.
In fact they created the curriculum for a new discipline: *Empirical Inference Science*.

(3) Developing high-quality professional software for empirical inference problems that can be downloaded and used by anyone in the world.[5]

This generation took advantage of computer technology to change forever the style and atmosphere of data mining research: from the very hierarchical group structure of the 1970–1980s lead by old statistical gurus (with their *know-how* and dominating opinion) to an open new society (with widely available information, free technical tools, and open professional discussions).

Many of the third generation researchers of SLT became university professors. This Afterword is dedicated to their students.

[5]The three most popular software are:
(1) SVM-*Light* developed by Thorsten Joachims (Germany) http://svmlight.joachims.org/,
(2) *Lib*-SVM developed by Chin-Chang Chang and Chih-Jen Lin (Taiwan) http://www.csie.ntu.edu.tw/~cjlin/libsvm/, and
(3) SVM-*Torch* developed by Ronan Collobert (Switzerland) http://www.torch.ch/

2.7 Relation to the Philosophy of Science

By the end of the 1990s it became clear that there were strong ties between machine learning research and research conducted in the classical philosophy of induction. The problem of generalization (induction) always was one of the central problems in philosophy. Pattern recognition can be considered as the simplest problem of generalization (its drosophila fly: any idea of generalization has its reflection in this model). It forms a very good object for analysis and verification of a general inductive principle. Such analysis includes not only speculations but also experiments on computers.

Two main principles of induction were introduced in classical philosophy: the principle of simplicity (parsimony) formulated by the 14th century English monk Occam (Ocham), and the principle of falsifiability, formulated by the Austrian philosopher of the 20th century Karl Popper. Both of them have a direct reflection in statistical learning theory.

2.7.1 Occam's Razor Principle

The Occam's Razor (or parsimony) principle was formulated as follows:

Entities are not to be multiplied beyond necessity.

Such a formulation leaves two open questions:

(1) What are the *entities*?

(2) What does *beyond necessity* mean?

According to *The Concise Oxford Dictionary of Current English* [172] the word *entity* means

> A thing's existence, as opposite to its qualities or relations; thing that has real existence.

So the number of entities is commonly understood to be the number of different parameters related to different physical (that which can be measured) features. The predictive rule is a function defined by these features.

The expression *not to be multiplied beyond necessity* has the following meaning: *not more than one needs to explain the observed facts.*

In accordance with such an interpretations the Occam's Razor principle can be reformulated as follows:

> *Find the function from the set with the smallest number of free parameters that explains the observed facts.*[6]

[6]There exist wide interpretation of Occam's Razor principle as a request to minimize some functional (without specifying which). Such interpretation is too general to be useful since it depends on the definition of the functional. The original Occam formulation (assuming that entities are free parameters) is unambiguous and in many cases is a useful instrument of inference.

2.7.2 Principles of Falsifiability

To introduce the principles of falsifiability we need some definitions.

Suppose we are given a set of indicator functions $f(x, \alpha), \alpha \in \Lambda$. We say that the set of vectors

$$x_1, \ldots, x_\ell, \; x_i \in X \tag{2.66}$$

cannot falsify the set of indicator functions $f(x, \alpha), \alpha \in \Lambda$ if all 2^ℓ possible separation of vectors (2.66) into two categories can be accomplished using functions from this set.

This means that on the data (2.66) one can obtain any classification (using functions from the admissible set). In other words, from these vectors one can obtain any possible law (given appropriate y_i, $i = 1, \ldots \ell$): the vectors themselves do not forbid (do not falsify) any possible law.

We say that the set of vectors (2.66) *falsifies* the set $f(x, \alpha), \alpha \in \Lambda$ if there exists such separation of the set (2.66) into two categories that cannot be obtained using an indicator function from the set $f(x, \alpha), \alpha \in \Lambda$.

Using the concept of falsifiability of a given set of functions by the given set of vectors, two different combinatorial definitions of the dimension of a given set of indicator functions were suggested: the VC dimension and the Popper dimension. These definitions lead to different concepts of falsifiability.

The Definition of the VC Dimension and VC Falsifiability

The VC dimension is defined as follows (in *EDBED* it is called capacity. See Chapter 6, Sections 8 and A2:)

A set of functions $f(x, \alpha), \alpha \in \Lambda$ has VC dimension h if:

(1) **there exist** h vectors that cannot falsify this set and

(2) **any** $h + 1$ vectors falsify it.

The set of functions $f(x, \alpha), \alpha \in \Lambda$ is *VC falsifiable* if its VC dimension is finite and *VC nonfalsifiable* if its VC dimension is infinite.

The VC dimension of the set of hyperplanes in R^n is $n + 1$ (the number of free parameters of a hyperplane in R^n) since there exist $n + 1$ vectors that cannot falsify this set but any $n + 2$ vectors falsify it.

The Definition of the Popper Dimension and Popper Falsifiability

The Popper dimension is defined as follows [137, Section 38]

A set of functions $f(x, \alpha), \alpha \in \Lambda$ has the Popper dimension h if:

(1) **any** h vectors cannot falsify it and

(2) **there exist** $h + 1$ vectors that can falsify this set.

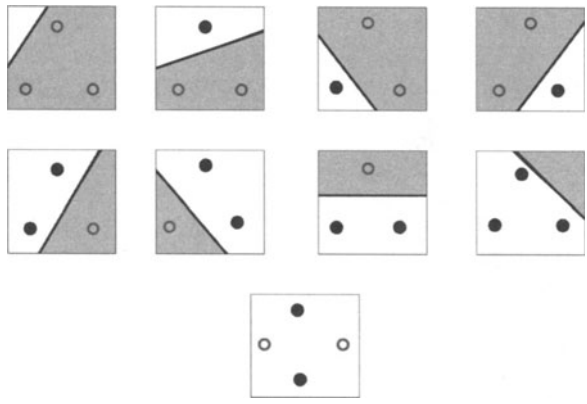

Figure 2.2: The VC dimension of the set of oriented lines in the plane is three since there exist three vectors that cannot falsify this set and any four vectors falsify it.

Popper called value h the degree of falsifiability or the dimension.

The set of functions $f(x, \alpha), \alpha \in \Lambda$ is *Popper falsifiable* if its Popper dimension is finite and *Popper nonfalsifiable* if its Popper dimension is infinite.

Popper's dimension of the set of hyperplanes in R^n is at most two (independent of the dimensionality of the space n) since only two vectors that belong to the one-dimensional linear manifold can not falsify the set of hyperplanes in R^n and three vectors from this manifold falsify this set.

2.7.3 Popper's Mistakes

In contrast to the VC dimension, the Popper concept of dimensionality does not lead to useful theoretical results for the pattern recognition model of generalization. The requirements of nonfalsifiability for any h vectors include, for example, the nonfalsifiability of vectors belonging to the line (one-dimensional manifold). Therefore, Popper's dimension will be defined by combinatorial properties restricted at most by the one-dimensional situation.

Discussing the concept of simplicity, Popper made several incorrect mathematical claims. This is the most crucial:

> In an algebraic representation, the dimension of a set of curves depends upon the number of parameters whose value we can freely choose. We can therefore say that the number of freely determinable parameters of a set of curves by which a theory is represented is characteristic of the degree of falsifiability. [137, Section 43]

This is wrong for the Popper dimension. The claim is correct only in a restricted situation for the VC dimension, namely when the set of functions in $R^n, n > 2$ *linearly depends on the parameters*.

In other (more interesting) situations as in Denker's example with a set of $\theta(\{\sin ax\})$ functions (Section 2.2.1 and Section 2.7.5 below) and in the example of a separating hyperplane with the margin given in *EDBED* (Chapter 10, Section 5) that led to SVM technology, the considered set of functions depends nonlinearly upon the free parameters.

Popper did not distinguish the type of dependency on the parameters. Therefore he claimed that the set $\{\theta(\sin ax)\}$ (with only one free parameter a) is a simple set of functions [137, Section 44]. However, the VC dimension of this set is infinite[7] and therefore generalization using this set of functions is impossible.

It is surprising that the mathematical correctness of Popper's claims has never been discussed in the literature.[8]

2.7.4 Principle of VC Falsifiability

In terms of the philosophy of science, the structural risk minimization principle for the structure organized by the nested set with increasing VC dimension can be reformulated as follows:

Explain the facts using the function from the set that is easiest to falsify.

The mathematical consistency of SRM therefore can have the following philosophical interpretation:

> Since one was able to find the function that separates the training data well, in the set of functions that is easy to falsify, these data are very special and the function which one chooses reflects the intrinsic properties of these data.[9].

It is possible, however, to organize the structure of nested elements on which capacity is defined by a more advanced measure than VC dimension (say, the Growth

[7]Since for any ℓ the set of values $x_1 = 2^{-1}, \ldots, x_\ell = 2^{-\ell}$ cannot falsify $\{\theta(\sin ax)\}$. The desired classifications $y_1, \ldots, y_\ell$, $y_i \in \{1, -1\}$ of this set provide the function $y = \theta(\sin a^* x)$ where the coefficient a^* is

$$a^* = \left(\pi \sum_{i=1}^{\ell} \frac{(1 - y_i)}{2} 2^i + 1 \right).$$

[8]Karl Popper's books were forbidden in the Soviet Union because of his criticism of communism. Therefore, I had no chance to learn about his philosophy until Gorbachev's time. In 1987 I attended a lecture on Popper's philosophy of science and learned about the falsifiability concept. After this lecture I became convinced that Popper described the VC dimension. (It was hard to imagine such a mistake.) Therefore in my 1995 and 1998 books I wrongly referred to Popper falsifiability as VC falsifiability. Only in the Spring of 2005 in the process of writing a philosophical article (see Corfield, Schölkopf, and Vapnik: "Popper, falsification and the VC dimension." Technical Report # 145, Max Planck Institute for Biological Cybernetics, Tübingen, 2005) did we check Popper's statements and realize my mistake.

[9]The *Minimum Message Length (MML)–Minimum Description Length* (MDL) principle [127, 128] that takes Kolmogorov's *algorithmic complexity* [129] into account can have the same interpretation. It is remarkable that even though the concepts of VC dimension and algorithmic complexity are very different, the MML-MDL principle leads to the same generalization bound for the pattern recognition problem that is given in *EDBED*. (See [139], Chapter 4, Section 4.6.)

function, or even better the VC entropy). This can lead to more advanced inference techniques (see Section 2.8 of this chapter).

Therefore the falsifiability principle is closely related to the VC dimension concept and can be improved by more refined capacity concepts.

2.7.5 Principle of Parsimony and VC Falsifiability

The principle of simplicity was introduced as a principle of parsimony or a principle of economy of thought.

The definition of simplicity, however, is crucial since it can be very different. Here is an example. Which set of functions is simpler:

(1) One that has the parametric form

$$f(x, \alpha), \ \alpha \in \Lambda, \ \text{ or}$$

(2) One that has the parametric form

$$f(x, \alpha), \ \alpha \in \Lambda$$

and satisfies the constraint

$$\Omega(f) \leq C,$$

where $\Omega(f) \geq 0$ is some functional?

From a computational point of view, finding the desired function in situation 1 can be much simpler than in situation 2 (especially if the $\Omega(f) \leq C$ is a nonconvex set).

From an information theory point of view, however, to find the solution in situation 2 is simpler, since one is looking for the solution in a more restricted set of functions.

Therefore the inductive principle based on the (intuitive) idea of simplicity can lead to a contradiction. That is why Popper used the "degree of falsifiability" concept (Popper dimension) to characterize the simplicity:

> The epistemological question which arise in connection with the concept of simplicity can all be answered if we equate this concept with degree of falsifiability. ([137], Section 43)

In the Occam's Razor principle, the number of "entities" defines the simplicity. Popper incorrectly claimed the equality of Popper dimension to be the number of free parameters (entities), and considered the falsifiability principle to be a justification of the parsimony (Occam's Razor) principle.

The principle of VC falsifiability does not coincide with the Occam's Razor principle of induction, and this principle (but not Occam's Razor) guarantee the generalization. VC dimension describes diversity of the set of functions. It does not refer either to the number of free parameters nor to our intuition of simplicity. Recall once again that Popper (and many other philosophers) had the intuition that $\{\theta(\sin ax)\}$ is the simple set of functions,[10] while the VC dimension of this set is infinite.

[10] In the beginning of Section 44 [137] Popper wrote: "According to common opinion the sine-function is a simple one . . .."

The principle of VC falsifiability forms the necessary and sufficient conditions of consistency for the pattern recognition problem while there are pattern recognition algorithms that contradict the parsimony principle.[11]

2.8 Inductive Inference Based on Contradictions

In my 1998 book, I discussed an idea of inference through contradictions [140, p.707]. In this Afterword, I introduce this idea as an algorithm for SVM. Sections 2.8.1 and 3.1.5 give the details of the algorithm. This section presents a simplified description of the general concept (see remark in Section 3.1.3 for details) of inductive inference through contradictions.

Suppose we are given a set of admissible indicator functions $f(x, \alpha), \alpha \in \Lambda$ and the training data. The vectors x from the training data split our admissible set of functions into a finite number of equivalence classes $F_1, \ldots, F_N$. The equivalence class contains functions that have the same values on the training vectors x (separate them in the same way).

Suppose we would like to make a structure on the set of equivalence classes to perform SRM principle. That is, we would like to collect some equivalence classes in the first element of the structure, then add to them some other equivalence classes, constructing the second element, and so on. To do this we need to characterize every equivalence class by some value that describes our preference for it. Using such a measure, one can create the desired structure on the equivalence classes. When we constructed SVMs, we characterized the equivalence class by the size of the largest margin defined by the hyperplane belonging to this class.

Now let us consider a different characteristic. Suppose along with the training data we possess a set of vectors called *the Universum* or *the Virtual Universum*

$$x_1^*, \ldots, x_k^*, \quad x^* \in X. \tag{2.67}$$

The Universum plays the role of prior information in Bayesian inference. It describes our knowledge of the problem we are solving. However, there are important differences between the prior information in Bayesian inference and the prior information given by the Universum. In Bayesian inference, prior information is information about the relationship of the functions in the set of admissible functions to the desired one. With the Universum, prior information is information related to possible training and test vectors. For example, in the digit recognition problem it can be some vectors whose

[11] The example of a machine learning algorithm that contradicts the parsimony principle is *boosting*. This algorithm constructs so-called weak features (entities) which it linearly combines in a decision rule. Often this algorithm constructs some set of weak features and the corresponding decision rule that separates the training data with no mistakes but continues to add new weak features (new entities) to construct a better rule. With an increasing number of (unnecessary, i.e., those that have no effect on separating the training data) weak features, the algorithm improves its performance on the test data. One can show that with an increasing number of entities this algorithm increases the margin (as the SVM). The idea of this algorithm is *to increase the number of entities (number of free parameters) in order to decrease the VC dimension* [136].

images resemble a particular digit (say some artificial characters). It defines the style of the digit recognition task, and geometrically belongs to the same part of input space to which the training data belong.

We use the Universum to characterize the equivalence class. We say that a vector x^* is contradictive for the equivalence class F_s if there exists a function $f_1(x^*) \in F_s$ such that

$$f_1(x^*) > 0$$

and there also exists a function $f_2(x^*) \in F_s$ such that

$$f_2(x^*) < 0.$$

We will characterize our preference for an equivalence class by the number of contradictions that occur on the Universum: the more contradictions, the more preferable the equivalence class.[12] We construct structure on equivalence classes using these numbers.

When using the Universum to solve a classification problem based on SRM principle, we choose the function (say one that has the maximal margin) from the equivalence class that makes no (or a small number of) training mistakes and has the maximal number of contradictions on the Universum. In other words, for inductive inference, when constructing the structure for SRM, we replace the *maximal margin* score with the *maximal contradiction on Universum* (MCU) score and select maximal margin function from the choosen equivalence class.

The main problem with MCU inference is, how does one create the appropriate Universum? Note that since one uses Universum only for evaluation of sizes of equivalence classes, its elements do not need to have the same distribution as the training vectors.

2.8.1 SVMs in the Universum Environment

The inference through contradictions can be implemented using SVM techniques as follows. Let us map both the training data and the Universum into Hilbert space

$$(y_1, z_1), \ldots, (y_\ell, z_\ell) \tag{2.68}$$

$$z_1^*, \ldots, z_u^*. \tag{2.69}$$

Quadratic Optimization Framework

In the quadratic optimization framework for an SVM, to conduct inference through contradictions means finding the hyperplane

$$(w^0, z) + b_0 = 0 \tag{2.70}$$

[12] A more interesting characteristic of an equivalence class would be the value of the VC entropy of the set of functions belonging to this equivalence class calculated on the Universum. This, however, leads to difficult computational problems. The number of contradictions can be seen as a characteristic of the entropy.

that minimizes the functional

$$R = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=1}^{u} \theta(\xi_j^*),\ C_1, C_2 > 0 \tag{2.71}$$

subject to the constraints

$$y_i((w, z_i) + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0,\ i = 1, \ldots, \ell \tag{2.72}$$

(related to the training data) and the constraints

$$|(w, z_j^*) + b| \leq a + \xi_j^*,\ \ \xi_j^* \geq 0,\ j = 1, \ldots, u \tag{2.73}$$

(related to the Universum) where $a \geq 0$.

As before, for computational reasons we approximate the target function (2.71) by the function[13]

$$R = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{s=1}^{u} \xi_s^*,\ C_1, C_2 > 0. \tag{2.74}$$

Using the Lagrange multipliers technique we determine that our hyperplane in feature space has the form

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i (z_i, z) + \sum_{s=1}^{u} (\mu_s^0 - \nu_s^0)(z_s^*, z) + b = 0, \tag{2.75}$$

where the coefficients $\alpha_i^0 \geq 0$, $\mu_s^0 \geq 0$, and $\nu_s^0 \geq 0$ are the solution of the following optimization problem: Maximize the functional

$$W(\alpha, \mu, \nu) = \sum_{i=1}^{\ell} \alpha_i - a \sum_{s=1}^{u} (\mu_s + \nu_s) - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (z_i, z_j) \tag{2.76}$$

$$- \sum_{i=1}^{\ell} \sum_{s=1}^{u} \alpha_i y_i (\mu_s - \nu_s)(z_i, z_s^*) - \frac{1}{2} \sum_{s,t=1}^{u} (\mu_s - \nu_s)(\mu_t - \nu_t)(z_s^*, z_t^*)$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i + \sum_{s=1}^{u} (\mu_s - \nu_s) = 0 \tag{2.77}$$

and the constraints

$$0 \leq \alpha_i \leq C_1 \tag{2.78}$$

$$0 \leq \mu_s, \nu_s \leq C_2. \tag{2.79}$$

[13] One also can use a least squares technique by choosing ξ_i^2 and $(\xi_i^*)^2$ instead of ξ_i and ξ_i^* in objective function (2.74).

Taking into account Mercer's theorem, one can rewrite our separating function in input space as

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x) + \sum_{s=1}^{u} (\mu_s^0 - \nu_s^0) K(x_s^*, x) + b_0 = 0, \tag{2.80}$$

where the coefficients $\alpha_i^0 \geq 0$, $\mu_s^0 \geq 0$, and $\nu_s^0 \geq 0$ are the solution of the following optimization problem: Maximize the functional

$$W(\alpha, \mu, \nu) = \sum_{i=1}^{\ell} \alpha_i - a \sum_{s=1}^{u} (\mu_s + \nu_s) - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.81}$$

$$- \sum_{i=1}^{\ell} \sum_{s=1}^{u} \alpha_i y_i (\mu_s - \nu_s) K(x_i, x_s^*) - \frac{1}{2} \sum_{s,t=1}^{u} (\mu_s - \nu_s)(\mu_t - \nu_t) K(x_s^*, x_t^*)$$

subject to the constraints (2.77), (2.78), (2.79).

Linear Optimization Framework

To conduct inference based only on contradictions arguments (taking some function from the choosen equivalence class, not necessarily one with the largest margin) one has to find the coefficients α^0, μ^0, ν^0 in (2.80) using the following linear programming technique: Minimize the functional

$$W(\alpha, \mu, \nu) = \gamma \sum_{i=1}^{\ell} \alpha_i + \gamma \sum_{s=1}^{u} (\mu_s + \nu_s) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{t=1}^{u} \xi_t^*, \ \gamma \geq 0 \tag{2.82}$$

subject to the constraints

$$y_i \left[\sum_{j=1}^{\ell} \alpha_j y_j K(x_i, x_j) + \sum_{s=1}^{u} (\mu_s - \nu_s) K(x_i, x_s^*) + b \right] \geq 1 - \xi_i, \ i = 1, \ldots, \ell \tag{2.83}$$

and the constraints

$$\sum_{j=1}^{\ell} \alpha_j y_j K(x_t^*, x_j) + \sum_{s=1}^{u} (\mu_s - \nu_s) K(x_t^*, x_s^*) + b \leq a + \xi_t^*, \ t = 1, \ldots, k, \tag{2.84}$$

$$\sum_{j=1}^{\ell} \alpha_j y_j K(x_t^*, x_j) + \sum_{s=1}^{u} (\mu_s - \nu_s) K(x_t^*, x_s^*) + b \geq -a - \xi_t^*, \ t = 1, \ldots, u, \tag{2.85}$$

where $a \geq 0$. In the functional (2.82) the parameter $\gamma \geq 0$ controls the sparsity of the solution.

2.8.2 The First Experiments and General Speculations

In the summer of 2005, Ronan Collobert and Jason Weston conducted the first experiments on training SVM with Universum using the algorithm described in Section 2.8.1. They discriminated digit 8 from digit 5 from the MNIST database, using a conventional SVM and an SVM trained in three different Universum environments.

The following table shows for different sizes of training sets the performance of a conventional SVM and the SVMs trained using Universums U_1, U_2, U_3 (each containing 5000 examples). In all cases the parameter $a = .01$, the parameters C_1, C_2, and the parameter of the Gaussian kernel were tuned using the tenfold cross-validation technique.

The Universums were constructed as follows:

U_1: Selects random digits from the other classes (0,1,2,3,4,6,7,9).

U_2: Creates an artificial image by first selecting a random 5 and a random 8, (from pool of 3,000 non-test examples) and then for each pixel of the artificial image choosing with probability 1/2 the corresponding pixel from the image 5 or from the image 8.

U_3: Creates an artificial image by first selecting a random 5 and a random 8, (from pool of 3,000 non-test examples) and then constructing the mean of these two digits.

No. of train. examples	250	500	1000	2000	3000
Test Err. SVM (%)	2.83	1.92	1.37	0.99	0.83
Test Err. SVM+U_1 (%)	2.43	1.58	1.11	0.75	0.63
Test Err. SVM+U_2 (%)	1.51	1.12	0.89	0.68	0.60
Test Err. SVM+U_3 (%)	1.33	0.89	0.72	0.60	0.58

The table shows that:
(a) The Universum can significantly improve the performance of SVMs.
(b) The role of knowledge provided by the Universum becomes more important with decreasing training size. However, even when the training size is large, the Universum still has a significant effect on performance.

We expect that advancing the understanding of the idea how to create a good Universum for the problem of interest will further boost the performance. This fact opens a new dimension in machine learning technology: How does one create a Virtual Universum for the problem of interest?

In trying to find an interpretation of the role of the Universum in machine learning, it is natural to compare it to the role of culture in the learning of humans, where knowledge about real life is concentrated not only in examples of reality but also in artificial images that reflect this reality. To classify well, one uses inspiration from both sources.

Chapter 3

NONINDUCTIVE METHODS OF INFERENCE: DIRECT INFERENCE INSTEAD OF GENERALIZATION (2000–· · ·)

3.1 INDUCTIVE AND TRANSDUCTIVE INFERENCE

Chapter 10 of *EDBED* distinguishes between two different problems of estimation: estimation of the function and estimation of the values of the function at given points of interest.

(1) *Estimation of the function.* Given training data

$$(y_1, x_1), \ldots, (y_\ell, x_\ell), \tag{3.1}$$

find in the set of admissible functions $f(x, \alpha), \alpha \in \Lambda$ the one which guarantees that its expected loss is close to the smallest loss.

(2) *Estimation of the value of the function at the points of interest.* Given a set of training data (3.1) and a sequence of k test vectors

$$x_{\ell+1}, \ldots, x_{\ell+k}, \tag{3.2}$$

find among an admissible set of binary vectors

$$\{Y_* = (y^*_{\ell+1}, \ldots, y^*_{\ell+k})\}$$

the one that classifies the test vectors with the smallest number of errors. Here we consider

$$x_1, \ldots, x_{\ell+k} \tag{3.3}$$

to be random i.i.d. vectors drawn according to the same (unknown) distribution $P(x)$. The classifications y of the vectors x are defined by some (unknown) conditional probability function $P(y|x)$.

This setting is quite general. In the book we considered a particular setting where the set of admissible vectors is defined by the admissible set of indicator functions $f(x, \alpha)$, $\alpha \in \Lambda$. In other words, every admissible vector of classification Y_* is defined as follows

$$Y_* = (f(x_1, \alpha_*), \ldots, f(x_k, \alpha_*)) .$$

In the mid-1990s (after understanding the relationship between the pattern recognition problem and the philosophy of inference), I changed the technical terminology [139]. That is, I called the problem of function estimation that requires one to find a function given particular data *inductive inference*. I called the problem of estimating the values of the function at particular points of interest given the observations *transductive inference*.

These two different ideas of inference reflect two different philosophies, which we will discuss next.

3.1.1 Transductive Inference and the Symmetrization Lemma

The mechanism that provides an advantage to the transductive mode of inference over the inductive mode was clear from the very beginning of statistical learning theory. It can be seen in the proof of the very basic theorems on uniform convergence. This proof is based on the following inequality which is the content of the so-called symmetrization lemma (see Basic lemma in *EDBED* Chapter 6, Section A3):

$$P\left\{\sup_{\alpha} |R(\alpha) - R_{emp}(\alpha)| \geq \varepsilon\right\} \leq 2P\left\{\sup_{\alpha} \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| \geq \frac{\varepsilon}{2}\right\}, \quad (3.4)$$

where

$$R^{(1)}_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - f(x_i, \alpha)| \quad (3.5)$$

and

$$R^{(2)}_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} |y_i - f(x_i, \alpha)| \quad (3.6)$$

are the empirical risks constructed using two different samples.

The bound for uniform convergence was obtained as an upper bound of the right-hand side of (3.4).

Therefore the symmetrization lemma implies that to obtain a bound for inductive inference we first obtain a bound for transductive inference (for the right-hand side of (3.4)) and then obtain an upper bound for it.

It should be noted that since the bound on uniform convergence was introduced in 1968, many efforts were made to improve it. However, all attempts maintain some form of the symmetrization lemma. That is, in the proofs of the bounds for uniform convergence the first (and most difficult) step was to obtain the bound for transductive inference. The trivial upper bound of this bound gives the desired result for inductive inference.

This means that transductive inference is a fundamental step in machine learning.

3.1.2 Structural Risk Minimization for Transductive Inference

The proof of the symmetrization lemma is based on the following observation: The following two models are equivalent (see Chapter 10, Section 1 of EDBED):

(a) one chooses two i.i.d. sets[1]

$$x_1, \ldots, x_\ell, \quad \text{and} \quad x_{\ell+1}, \ldots, x_{2\ell};$$

(b) one chooses an i.i.d. set of size 2ℓ and then randomly splits it into two subsets of size ℓ.

Using model (b) one can rewrite the right-hand side of (3.4) as follows

$$P\left\{\sup_\alpha \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| > \frac{\varepsilon}{2}\right\} =$$

$$E_{\{x_1,\ldots,x_{2\ell}\}} P\left\{\sup_\alpha |R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)| > \frac{\varepsilon}{2} \; |\{x_1, \ldots, x_{2\ell}\}\right\}. \tag{3.7}$$

To obtain the bound we first bound the conditional probability

$$P\left\{\sup_\alpha |R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)| > \frac{\varepsilon}{2} \; |\{x_1, \ldots, x_{2\ell}\}\right\} \leq$$

$$\Delta^\Lambda(x_1, \ldots, x_{2\ell}) \exp\left\{-\varepsilon^2 \ell\right\} \tag{3.8}$$

where $\Delta^\Lambda(x_1, ..., x_{2\ell})$ is the number of equivalence classes on the set (3.3). The probability is obtained with respect to the random split data into two parts (training and testing). Then we take the expectation over working sets of size 2ℓ. As a result, we obtain

$$E_{\{x_1,\ldots,x_{2\ell}\}} P\left\{\sup_\alpha \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| > \frac{\varepsilon}{2} |\{x_1, \ldots, x_{2\ell}\}\right\} \leq$$

$$\Delta^\Lambda_P(2\ell) \exp\left\{-\varepsilon^2 \ell\right\}. \tag{3.9}$$

Note that for the transductive model of inference we do not even need to take the expectation over sets of size 2ℓ. We can just use the bounds (3.8).

[1] For simplicity of the formulas we choose two sets of equal size.

Let us consider both models of inference, transductive and inductive from one unified point of view: In both cases we are given a set of functions defined on some space R. We randomly choose the training examples from this space. In the inductive case we choose by sampling from the space, and in the transductive case we choose by splitting the working set into the training and testing parts. We define the values of the function of interest over the domain of definition of the function: In the inductive case in the whole space, and in the transductive case on the working set.

The difference is that in transductive inference the space of interest is discrete (defined on $\ell + k$ elements of the working set (3.3)), while in inductive inference it is R^n.

One can conduct a nontrivial analysis of the discrete space but not the continuous space R^n. This is the key advantage of transductive inference.

3.1.3 Large Margin Transductive Inference

Let $F_1, \ldots, F_N$ be the set of equivalence classes defined by the working set (3.3). Our goal is to construct an appropriate structure on this set of equivalence classes.

In Chapter 2, Section 2.6 we constructed a similar structure on the set of equivalence classes for inductive inference. However, we violated one of the important requirements of the theory: The structure must be constructed *before* the training data appear. In fact we constructed it *after* (in the inductive mode of inference the set of equivalence classes was defined by the training data), creating a *data-dependent structure*. There are technical means to justify such an approach. However, the bound for a data-dependent structure will be worse [138].

In transductive inference we construct the set of equivalence classes using a joint working set of vectors that contain both the training and test sets. Since in constructing the equivalence class we do not use information about how our space will be split into training and test sets we do not violate the statistical requirements.

Let us define the size of an equivalence class F_i by the value of the corresponding margin: We find, among the functions belonging to the equivalence class, the one that has the largest margin[2] and use the value of the margin $\mu(F_i)$ as the size of the equivalence class F_i.

Using this concept of the size of an equivalence class, SVM transductive inference suggests:

Classify the test vectors (3.2) by the equivalence class (defined on the working set (3.3)) that classifies the training data well and has the largest value of the (soft) margin.

Formally, this requires us to classify the test data using the rule

$$y_i = \operatorname{sgn}((w_0, z_i) + b_0), \; i = \ell + 1, \ldots, \ell + k,$$

[2]We consider the hard margin setting just for the sake of simplicity. One can easily generalize this setting to the soft margin situation as described in Section 2.3.4.

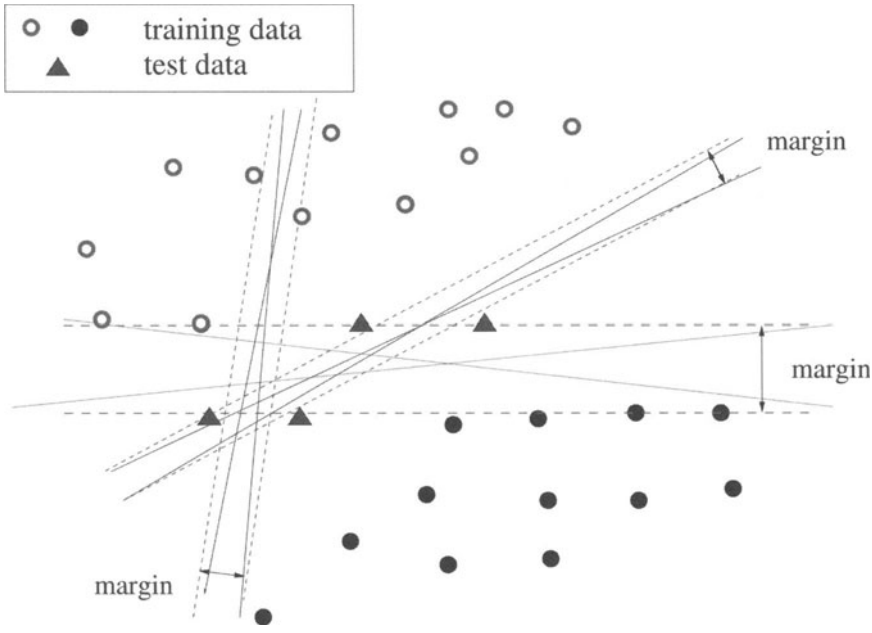

Figure 3.1: Large margin defines a large equivalence class.

where the parameters w_0, b_0 are the ones that minimize the functional

$$R(w) = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=\ell+1}^{\ell+k} \theta(\xi_j),\ C_1, C_2 \geq 0 \tag{3.10}$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0,\ i = 1, \ldots, \ell \tag{3.11}$$

(defined by the images of the training data (3.1)) and the constraints

$$y_j^*((z_j, w) + b) \geq 1 - \xi_j, \quad \xi_j \geq 0,\ j = \ell + 1, \ldots, \ell + k \tag{3.12}$$

(defined by the set (3.2) and its desired classification $Y_* = (y_{\ell+1}^*, \ldots, y_{\ell+k}^*)$.

One more constraint. To avoid unbalanced solution Chapelle and Zien [174], following ideas of Joachims [154], suggested the following constraint:

$$\frac{1}{k} \sum_{j=\ell+1}^{\ell+k} ((w, z_j) + b) \approx \frac{1}{\ell} \sum_{i=1}^{\ell} y_i. \tag{3.13}$$

This constraint requires that the proportion of test vectors in the first and second categories be similar to the proportion observed in the training vectors.

For computational reasons we will replace the objective function (3.10) with the function

$$R(w) = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \xi_i, + C_2 \sum_{s=\ell+1}^{\ell+k} \xi_s^*, \; C_1, C_2 \geq 0 \tag{3.14}$$

Therefore (taking into account kernelization based on Mercer's theorem) we can obtain the following solution of this problem (in the dual space).

The classification rules for the test data in the dual space have the form

$$y_\tau = \text{sgn}(\sum_{i=1}^{\ell} \alpha_i^0 K(x_i, x_\tau) + \sum_{s=\ell+1}^{\ell+k} \beta_s y_s^* K(x_s, x) + b_0), \; \tau = \ell+1, \ldots, \ell+k,$$

where the coefficients $\alpha_i^0, \beta_s^0, b_0$ and desired classifications of test data are the solution of the following problem: Maximize (over α, β, y^*) the functional

$$W(\alpha, \beta, y^*) = \sum_{i=1}^{\ell} \alpha_i + \sum_{s=\ell+1}^{\ell+k} \beta_s - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$- \sum_{i=1}^{\ell} \sum_{s=\ell+1}^{\ell+k} \alpha_i y_i \beta_s y_s^* K(x_i, x_s) - \frac{1}{2} \sum_{s,t=\ell+1}^{\ell+k} \beta_s y_s^* \beta_t y_t^* K(x_s, x_t)$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i + \sum_{s=\ell+1}^{\ell+k} y_s^* \beta_s = 0,$$

the constraints

$$0 \leq \alpha_i \leq C_1, \; i = 1, \ldots, \ell$$

$$0 \leq \beta_s \leq C_2, \; s = \ell+1, \ldots, \ell+k,$$

and the constraint (3.13):

$$\frac{1}{k} \sum_{j=\ell+1}^{\ell+k} \left(\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x_j) + \sum_{t=\ell+1}^{\ell+k} \beta_t y_t^* K(x_t, x_j b_0 \right) \approx \frac{1}{\ell} \sum_{i=1}^{\ell} y_i.$$

Note that this problem does not have a unique solution. This makes transductive inference difficult. However, whenever one can maximize the functional well, one obtains an improvement over inductive SVMs.

3.1.4 Examples of Transductive Inference

Here are examples of real-life problems solved using transductive inference.

1. Prediction of molecular bioactivity for drug discovery [146]. The KDD CUP-2001 competition on data analysis methods required the construction

of a rule for predicting molecular bioactivity using data provided by the DuPont Pharmaceutical company. The data belonged to a binary space of dimension 139,351, which contained a training set of 1909 vectors, and a test set of 634 vectors.

The results are given here for the winner of the competition (among the 119 competitors who used traditional approaches), SVM inductive inference and SVM transductive inference.

Winner's accuracy	68.1 %
SVM inductive mode accuracy	74.5 %
SVM transductive mode accuracy	82.3 %

It is remarkable that the jump in performance obtained due to a new philosophy of inference (transductive instead of inductive) was larger than the jump resulting from the reinforcement of the technology in the construction of inductive predictive rules.

2. TEXT CATEGORIZATION [138]. In a text categorization problem, using transductive inference instead of inductive inference reduced the error rate from 30% to 15%.

REMARK. The discovery of transductive inference and its advantages over inductive inference is not just a technical achievement, but a breakthrough in the philosophy of generalization.

Until now, the traditional method of inference was the *inductive–deductive* method, where one first defines a general rule using the available information, and then deduces the answer using this rule. That is, one goes from *particular to general* and then from *general to particular.*

In transductive mode one provides direct inference from *particular to particular*, avoiding the ill-posed part of the inference problem (inference from particular to general).

3.1.5 TRANSDUCTIVE INFERENCE THROUGH CONTRADICTIONS

Replacing the maximal margin generalization principle with the maximal contradiction on the Universum (MCU) princple leads to the following algorithm: Using the working set (3.3) create a set of equivalence classes of functions, then using the Universum (2.67) calculate the size of the equivalence classes by the number of contradictions.

The recommendation of SRM for such a structure would be:

> *To classify test vectors (3.2), choose the equivalence class (defined on the working set (3.3)) that classifies the training data (3.1) well and has the largest number of contradictions on the Universum.*

The idea of maximizing the number of contradictions on the Universum can have the following interpretation:

> *When classifying the test vectors, be very specific; try to avoid extra generalizations on the Universum (2.67).*

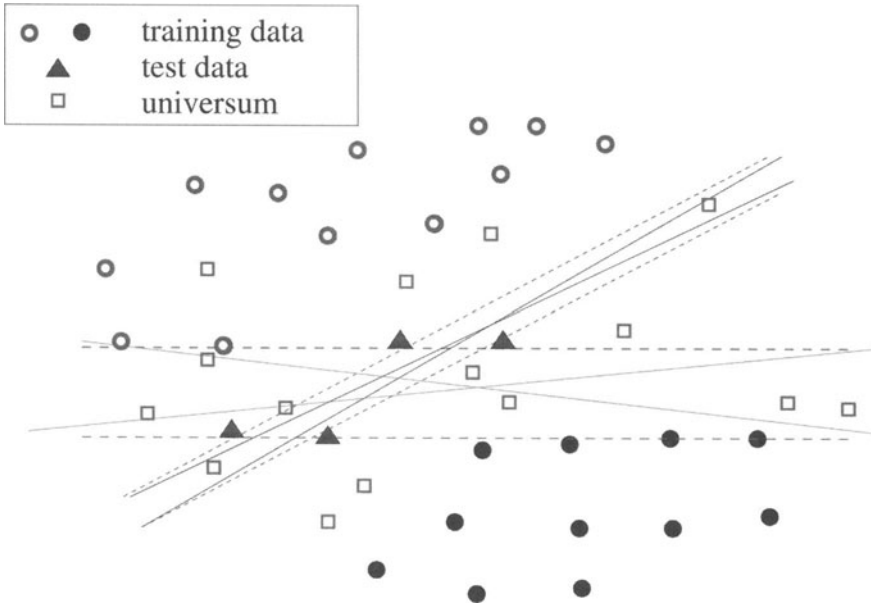

Figure 3.2: Large number of contradictions on Universum (boxes inside the margin) defines a large equivalence class.

From a technical point of view, the number of contradictions takes into account the inhomogeneity of image space, especially when the input vectors are nonlinearly mapped into feature space.

Technically, to implement transductive inference through contradictions one has to solve the following problem.

Given the images of the training data (3.1), the images of the test data (3.2), and the images of the Universum (2.67), construct the linear decision rule

$$I(x) = \theta[(w_0, z) + b_0],$$

where the vector w_0 and threshold b_0 are the solution of the following optimization problem: Minimize the functional

$$R(w) = \frac{1}{2}(w, w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=\ell+1}^{\ell+k} \theta(\xi_j) + C_3 \sum_{s=1}^{u} \theta(\xi_s^*), \quad C_1, C_2, C_3 \geq 0 \tag{3.15}$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0, \; i = 1, \ldots, \ell \tag{3.16}$$

(defined by the images of the training data (3.1)), the constraints

$$y_j^*((z_j, w) + b) \geq 1 - \xi_j, \quad \xi_j \geq 0, \; j = \ell + 1, \ldots, \ell + k \tag{3.17}$$

(defined by the set (3.2) and the desired vector $(y^*_{\ell+1}, \ldots, y^*_{\ell+k})$), and the constraints

$$|(z^*_s, w) + b| \le a + \xi^*_s, \quad \xi^*_s \ge 0, \ s = 1, \ldots, u, \ \ a \ge 0 \tag{3.18}$$

(defined by the images of the Universum (2.67)).

As before (for computational reasons), we replace $\theta(\xi)$ in the objective function with ξ. Therefore we minimize the functional

$$R(w) = \frac{1}{2}(w, w) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{j=\ell+1}^{\ell+k} \xi_j + C_3 \sum_{s=1}^{u} \xi^*_s, \quad C_1, C_2, C_3 \ge 0 \tag{3.19}$$

subject to the constraints (3.16), (3.17), and (3.18).

Dual Form Solution

The solutions to all of the above problems in the dual space of Lagrange multipliers can be unified as follows. Find the function

$$f(x) = \sum_{i=1}^{\ell} \alpha^0_i y_i K(x, x_i) + \sum_{t=\ell+1}^{\ell+k} \beta^0_t y^*_t K(x, x_t) + \sum_{m=1}^{u} (\mu^0_m - \nu^0_m) K(x, x^*_m) + b_0 \tag{3.20}$$

whose test classifications y^*_j and coefficients $\alpha^0, \beta^0, \mu^0, \nu^0, b_0$ maximise the functional

$$W(\alpha, \beta, \gamma, \mu, \nu, y^*) = \sum_{i=1}^{\ell} \alpha_i + \sum_{t=\ell+1}^{\ell+k} \beta_t - a \sum_{n=1}^{u} (\mu_n + \nu_n) \tag{3.21}$$

$$-\frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \frac{1}{2} \sum_{s,t=\ell+1}^{\ell+k} \beta_t y^*_t \beta_s y^*_s K(x_t, x_s)$$

$$-\frac{1}{2} \sum_{m,n=1}^{u} (\mu_m - \nu_m)(\mu_n - \nu_n) K(x^*_m, x^*_n) - \sum_{i=1}^{\ell} \sum_{t=\ell+1}^{\ell+k} \alpha_i y_i \beta_t y^*_t K(x_i, x_t)$$

$$-\sum_{i=1}^{\ell} \sum_{m=1}^{u} \alpha_i y_i (\mu_m - \nu_m) K(x_i, x^*_m) - \sum_{m=1}^{u} \sum_{t=\ell+1}^{\ell+k} (\mu_m - \nu_m) \beta_t y^*_t K(x^*_m, x_t)$$

subject to the constraints

$$0 \le \alpha_i \le C_1, \tag{3.22}$$

$$0 \le \beta_t, \le C_2, \tag{3.23}$$

$$0 \le \mu_m, \nu_m \le C_3, \tag{3.24}$$

and the constraint

$$\sum_{i=1}^{\ell} \alpha_i y_i + \sum_{t=\ell+1}^{\ell+k} \beta_t y^*_t + \sum_{m=1}^{u} (\mu_m - \nu_m) = 0. \tag{3.25}$$

In particular, when $C_2 = C_3 = 0$ we obtain the solution for the conventional SVM, when $C_2 = 0$ we obtain the solution for inductive SVMs with the Universum, and when $C_3 = 0$ we obtain the solution for transductive SVMs.

Note that just taking into account the Universum ($C_2 = 0$) does not change the convexity of the optimization task. The problem becomes nonconvex (and therefore can have a nonunique solution) only for transductive mode.

It is good to use hint (3.13) when solving transductive problems.

3.2 Beyond Transduction: The Transductive Selection Problem

The transductive selection problem was not discussed in the original Russian edition of *EDBED*. It was written at the last moment for the English translation. In *EDBED* the corresponding section (Chapter 10, Section 13) has a very technical title "The Problem of Finding the Best Point of a Given Set." Here we call this type of inference transductive selection.

3.2.1 Formulation of Transductive Selection Problem

The transductive selection problem is the following: Given the training examples (pairs $(x_i, y_i), x \in R^n$, $y \in \{-1, +1\}$, $i = 1, \ldots, \ell$) and given a working set ($x_j^* \in R^*$, $j = 1, \ldots, m$), find in the working set the k elements that belong to the first class ($y = +1$) with the highest probability.

Here are some examples of the selection problem:

- *Drug discovery.* In this problem, we are given examples of effective drugs $(x_i, +1)$ and examples of ineffective drugs $(x_s, -1)$. The goal is to find among the given candidates $(x_1^*, \ldots, x_m^*)$ the k candidates with the highest probability of being effective drugs.

- *National security.* In this problem, we are given examples (descriptions) of terrorists $(x_i, +1)$ and examples of non-terrorists $(x_s, -1)$. The goal is to find among the given candidates $(x_1^*, \ldots, x_m^*)$ the k most likely terrorists.

Note that in contrast to general transductive inference, this setting does not require the classification of all candidates[3]. The key to solving the selective inference problem is to create an appropriate factorization of a given set of functions that contains fewer equivalence classes than the factorization for transductive inference. The transductive selective models are the main instrument for solving decision-making problems in high-dimensional spaces. However, this instrument has not yet been developed.

[3]In such problems, the most difficult cases are "border candidates." In transductive selection problems, we exclude this most difficult part of the task (classification of border candidates). Here again we obtain the same advantage that we obtained by replacing the model identification scheme by the prediction scheme and replacing the predictive scheme by the transductive scheme: we replaced a not very well-posed problem by a better-posed problem.

3.3 Directed Ad Hoc Inference (DAHI)

3.3.1 The Idea Behind DAHI

This section discusses *directed ad hoc inference*, inference that occupies an intermediate position between *inductive–deductive* inference and *transductive* inference.

The main idea of DAHI is a reconsideration of the roles of the training and testing stages during the inference process. The classical *inductive–deductive* model of inference contains two different stages:

(1) The training (inductive) stage, where one constructs a rule for classification using the training data, and

(2) The testing (deductive) stage where one classifies the test data using the constructed rule.

The *transductive* model of inference solves the classification problem in one step:

- Given a set of training data and a set of test data, it finds the labels for the test data directly.

DAHI works differently. During the training stage, DAHI looks for a principal direction (concept) used to construct different rules for future inferences. This is different from the inductive stage of inference where the goal is to find one fixed rule. During the test stage DAHI uses this principal direction to construct a specific rule for each given test vector (the ad hoc rule). Therefore, DAHI contains elements of both inductive and transductive inference:

(1) It constructs one general direction of inference (as in inductive inference).

(2) It constructs an individual (ad hoc) rule for each given test example (as in transductive inference).

The idea of DAHI is: *To construct a linear (in feature space) decision rule that has fixed homogeneous terms and individual (for different test vectors) thresholds.*

The problem is how to find thresholds that make inferences more accurate than ones based on one fixed threshold (as in SVM).

From a technical point of view DAHI is a combination of ideas from statistical learning theory (in particular, support vector machines), and from nonparametric statistics (methods for conditional probability estimation).

3.3.2 Local and Semi-Local Rules

To discuss the details of DAHI let us consider the idea of *local algorithms* suggested by nonparametric statistics and in particular the k-nearest neighbors method.

k-NEAREST NEIGHBORS METHOD

According to the k-nearest neighbours method for any point of interest x_0 one chooses from the training data the k-nearest (in a given metric) vectors x_i, $i = 1, \ldots, k$ and classifies the point of interest x_0 depending on which class dominates among these k chosen vectors.

The k-nearest neighbors method can be described as a *local* estimating method. Consider the set of constant-valued functions. For a set of indicator functions it contains only two functions: one takes the value -1; another takes the value 1. Consider the following local algorithm: define the spherical vicinity of the point of interest x_0 based on the given metric and a value for the radius (defined by the distance from a point of interest x_0 to its k nearest neighbors). Then choose from the admissible set of functions the function that minimizes the empirical loss on the training vectors belonging to the vicinity of the point of interest x_0. Finally use this function to classify the point of interest.

This description of the k-nearest neighbors method as a local algorithm immediately allows one to generalize it in two respects:

(1) One can use a richer set of admissible functions (for example, the set of large margin linear decision rules, see Section 2.3)

(2) One can use different rules to specify the value of the radius that defines the locality (not just the distance to the kth nearest neighbor).

In 1992 the idea of local algorithms for pattern recognition was used where (local) linear rules (instead of local constant rules) and VC bounds (instead of the distance to the kth nearest neighbor) were utilized [145]. The local linear rules demonstrated a significant improvement in performance (3.2% error rate instead of 4.1% for digit recognition on the US Postal Service database).

For the regression estimation problem a similar idea was used in the Nadaraya–Watson estimator [147, 148] with a slightly different concept of locality. Nadaraya and Watson suggested considering "soft locality": they introduced a weight function (e.g., a monotonically decreasing nonnegative function from the distance between a point of interest x_0 and elements x_i of training data $f(||x_0 - x_i||)$, $i = 1, \ldots, \ell$), and used this function for estimating the value of interest

$$y_0 = \sum_{i=1}^{\ell} \tau_i(x_0) y_i, \tag{3.26}$$

where coefficients $\tau_i(x_0)$ were defined as follows,

$$\tau_i(x_0) = \frac{f(||x_0 - x_i||)}{\sum_{i=1}^{\ell} f(||x_0 - x_i||)}. \tag{3.27}$$

This concept is a generalization of the hard locality concept. We will use this construction later.

However in all of these methods the concept of locality is the same: it is a sphere (a "soft sphere" in the Nadaraya–Watson method) defined by a given metric with the center at the point of interest.

SEMI-LOCAL RULE

In DAHI we use a new concept of vicinity. We map input vectors x into a feature space z where we specify the vicinity. We consider a cylinder (or more generally a "soft cylinder"; see Section 3.3.4 below) whose axis passes through the image z_0 of the point of interest x_0. The defined vicinity is unbounded in one direction (defined by the axis of the cylinder) and bounded in all other directions. We call such a vicinity a *semi-local* vicinity.

The difference between the local and semi-local concepts of vicinity is the following. In a sphere with a fixed center there are no preferable directions in a feature space, while a cylinder has one preferable direction (along the axis of the cylinder). DAHI uses this direction to define vicinities for all points of interest.

During the training stage DAHI looks for the direction of the cylinder that defines the axis (in feature space) for all possible vicinities (cylinders). To find this direction one can use the methods of statistical learning theory (e.g., SVMs).

During the test stage DAHI uses only data from the (semi-local) vicinity of the point of interest z_0 and constructs a one-dimensional conditional probability function defined on the axis of the cylinder passing in the specified direction w_0 through the point of interest z_0. DAHI then uses this conditional probability $P(y_0 = 1|z_0)$ to classify z_0, where z_0 is the image of the point of interest x_0 in feature space.

Note that DAHI generalizes the SVM idea. In SVM one chooses both the direction w_0 and the threshold b_0 for the decision rule. In DAHI one chooses only the direction w_0, and for any test vector constructs an individual decision rule (threshold).

3.3.3 ESTIMATION OF CONDITIONAL PROBABILITY ALONG THE LINE

To solve the classification part of the problem we estimate the conditional probability $P(y(t) = 1|t)$ that the point t on the axis of a cylinder (passing through the point of interest t_0) belongs to the first class. To do this we have to solve the integral equation

$$\int_a^t P(y = 1|t')dF(t') = F(y = 1, t), \tag{3.28}$$

where both the cumulative distribution function of the point on the line $F(t)$ and the probability function $F(y = 1, t)$ of that point on the line with $t' \leq t$ belong to the first class are unknown, but data (inside cylinder) are given.

Note that when the density function $p(t)$ exists for $F(t)$, the conditional probability

$$P(y = 1|t) = \frac{p(y = 1, t)}{p(t)}$$

defines the solution of Equation (3.28).

To solve this problem given data one must first estimate the cumulative distribution functions along the line and then use these estimates $F_{est}(t)$, $F_{est}(1,t)$ in Equation (3.28) instead of the actual functions $F(\xi)$ and $F(y=1,\xi)$.

$$\int_a^t P(y=1|t')dF_{mp}(t') = F_{emp}(y=1,t). \tag{3.29}$$

This Equation forms an ill-posed problem where not only the right-hand side of the equation is an approximation of the real right-hand side but also the operator is an approximation of the real operator (since we use $F_{emp}(t)$ instead of $F(t)$).

In [140] it is shown that if the approximations $F_{emp}(t)$, and $F_{emp}(y=1,t)$ are consistent then there exists a law $\gamma_\ell = \gamma(\ell)$ such that the Tikhonov regularization method

$$R(P) = \left|\left| \int_a^t P(y=1|t')dF_{emp}(t') - F_{emp}(y=1,t) \right|\right|^2 + \gamma_\ell \Omega(P) \tag{3.30}$$

provides the solutions that converge to the solution of Equation (3.28) as $\ell \to \infty$.

3.3.4 Estimation of Cumulative Distribution Functions

A consistent method of estimating cumulative distribution functions along a line was first suggested by Stute in 1986 [149]. He considered a cylinder of radius r whose axis coincides with the line, projected on this line the vectors z of the training data that were inside the cylinder (suppose that there are $r(\ell)$ such vectors), and constructed a one-dimensional empirical distribution function using these projections:

$$F^*_{r(\ell)}(x) = \frac{1}{r(\ell)} \sum_{i=1}^{r(\ell)} \theta(t - t_i). \tag{3.31}$$

Stute showed that under some general law of choosing the radius of the cylinder (which depends on the number of observations ℓ) with an increasing number of observations, this empirical cumulative distribution function converges with probability one to the desired function. To estimate conditional probability one can use in (3.30) the approximation (3.31) and the approximation

$$F_{r(\ell)}(1,t) = \frac{1}{2r(\ell)} \sum_{i=1}^{r(\ell)} (1+y_i)\theta(t - t_i). \tag{3.32}$$

Also one can estimate a cumulative distribution function along the line in the Nadaraya-Watson style using the distances between images of training vectors and the line passing through the point of interest z_0 in direction w_0 ,

$$d_i(z_0) = \sqrt{|z_i - z_0|^2 - t_0^2}, \tag{3.33}$$

where $t_0 = (z_0, w_0)$ is the projection of the vector z_0 on the direction w_0. Using $d_i(z_0)$ instead of $||x_0 - x_i||$ in (3.27) one obtains the Nadaraya–Watson type approximations of the elements of Equation (3.29):

$$F_{emp}(t) = \sum_{i=1}^{\ell} \tau_i(z_0)\theta(t - t_i), \tag{3.34}$$

$$F_{emp}(y = 1, t) = \frac{1}{2}\sum_{i=1}^{\ell} (1 + y_i)\tau_i(z_0)\theta(t - t_i). \tag{3.35}$$

Both the Stute estimate and modified Nadaraya–Watson estimate are step functions. The difference is that in Stute's estimate there are $r(\ell)$ steps where all values of the step are equal to $1/r(\ell)$ while in the Nadarya–Watson estimate there are ℓ steps but the step values $\tau_i(z_0)$ are different, and depend on the distance between the vector z_i and the line passing through the point z_0 in the direction w_0.

3.3.5 Synergy Between Inductive and Ad Hoc Rules

In DAHI we combine two consistent methods: the SVM method for estimating the direction in feature space, and the method for estimating the conditional probability along the line passing through the point of interest.

However, when the number of training data is not large (and this is always the case in a high-dimensional problem) one needs to provide both methods with additional information: In order to choose a good SVM solution one has to map the input vectors into a "good" Hilbert space (to choose a "good" kernel). In order to obtain a good solution for solving the ill-posed problem of estimating a conditional probability function along the line one has to use a priori information about the admissible set of functions that contain the desired conditional probability function.

By combining the above two methods, one tries to construct a robust classification method that reduces the dependency on a priori information.

This is because:

(1) When one chooses a direction that is "reasonably close" to the one that defines a "good" separating hyperplane, the corresponding conditional probability function belongs to the set of *monotonic* functions (the larger the SVM score is, the larger is the probability of the positive class). Finding a direction that maintains the monotonicity property for the conditional probabilities requires fewer training examples than finding a direction that provides a good classification.

(2) The problem of finding a conditional probability function from the set of monotonic nondecreasing functions is much better posed than the more general problem of finding a solution from the set of continuous nonnegative functions.[4]

Therefore, in the set of monotonic functions one can solve this problem well, using a restricted (small) number of observations.

[4] A set of monotonically increasing (or monotonically decreasing) functions has VC dimension one while a set of continuous nonnegative functions has an infinite VC dimension.

(3) Using the leave-one-out technique one can use the same training data for constructing the main direction and later for constructing conditional probability functions.

The minimization of functional (3.30) in a set of monotonic functions is not too difficult a computational problem. The idea behind DAHI is to use this possible synergy.

Figure 3.3 shows two examples of the binary classification problem: separating digit 3 from digit 5. Two examples of conditional probabilities $P(3|t)$ estimated along the line are presented in Figure 3.3. For each example the figure shows the image of interest, the functions $F_{emp}(t)$ and $F_{emp}(3,t)$, and the solution of Equation (3.29). The position of the point of interest on the line corresponds to an ordinate value of 0. Part (a) of the figure shows the probability that the image is a 3 is 0.34, but in part (b) the probability that the image is a 3 is 0.

3.3.6 DAHI AND THE PROBLEM OF EXPLAINABILITY

The idea of DAHI is appealing from a philosophical point of view since it addresses the question of *explainability* of complex rules [169]. DAHI divides the model of explainability for complex rules into two parts: the "main direction" and the "ad hoc" parts where only the "main direction" part of the rule has to be explained (described by the formal model).

One speculation on the DAHI model of explainability can be given by the example of how medical doctors distinguish between cancer and benign cases. They use principal rules to evaluate the cancer and if the corresponding score exceeds a threshold value, they decide the case is cancer.

The threshold, however, is very individual: it depends on the family history of the patient, and many other factors. The success of a doctor depends on his experience in determining the individual threshold. The threshold can make all the difference in diagnostics. Nevertheless the explainability is mostly related to the "main direction" part of the rule.

3.4 PHILOSOPHY OF SCIENCE FOR A COMPLEX WORLD

3.4.1 EXISTENCE OF DIFFERENT MODELS OF SCIENCE

The limitations of the classical model of science when dealing with the real-life complex world have been discussed for quite some time. For example, according to Einstein, the classical model of science is relevant for a simple world. For a complex world it is inapplicable.

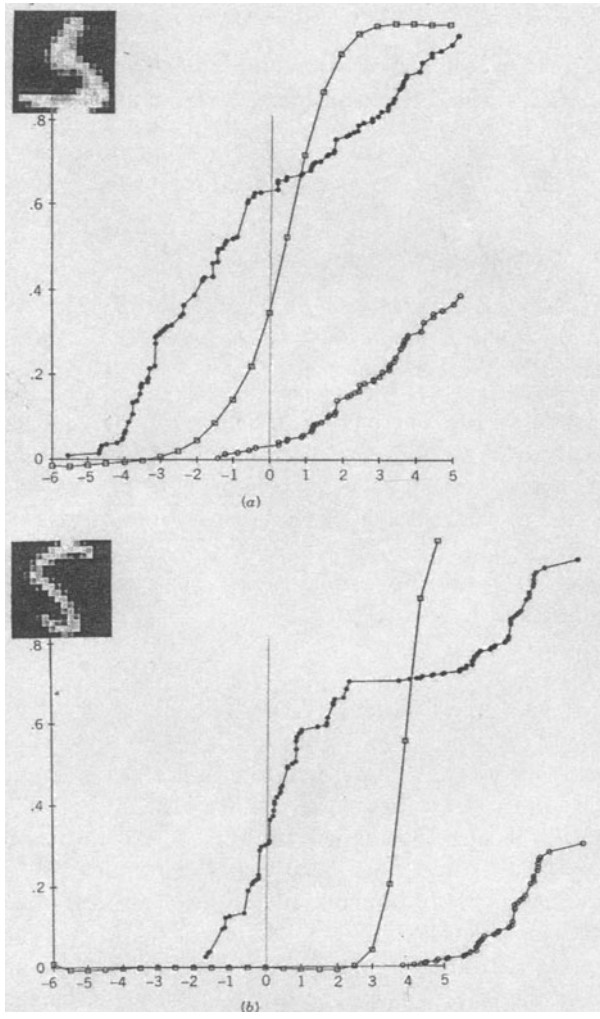

Figure 3.3: Solutions of the integral equation for different data.

- Einstein on the simple world:

  When the solution is simple, God[5] *is answering.*

- Einstein on the complex world:

  When the number of factors coming into play in a phenomenological complex is too large, scientific methods in most cases fail.[6]

One can see the idea of limitation of scientific models and existence of non-scientific ones in the following Richard Feynman's remark (*Lectures on physics*):

> *If something is said not to be a science, it does not mean that there is something wrong with it ... it just means that it is not a science.*

In other words there was an understanding that:

> *Classical science is an instrument for a simple world. When a world is complex, in most cases classical science fails. For a complex world there are methods that do not belong to classical science.*

Nevertheless, the success of the physical sciences strongly influenced the methodology used to analyze the phenomena of a complex world (one based on many factors). In particular, such a methodology was adopted in the biological, behavioural, and social sciences where researchers tried to construct low-dimensional models to explain complex phenomena.

The development of machine learning technology challenged the research in the methodology of science.

3.4.2 Imperative for a Complex World

Statistical learning theory stresses that the main difficulties of solving generalization problems arise because, in most cases, they are ill-posed.

To be successful in such situations, it suggests to give up attempts of solving ill-posed problems of interest replacing them by less demanding but better posed problems. In many cases this leads to renunciation of explainability of obtained solutions (which is one of the main goals declared by the classical science). Therefore, a science for a complex world has different goals (may be it should be called differently).

For solving specific ill-posed problems the regularization technique was suggested [20, 21, 54, 55]. However, to advance high-dimensional problems of inference just applying classical regularization ideas is not enough. The SRM principle of inference is another way to control the capacity of admissible sets of functions. Recently a new general idea of capacity control was suggested in the form of the following imperative [139]:

[5] Here and below Einstein uses the word God as a metaphor for nature.

[6] Great theoretical physicist Lev Landau did not trust physical theories that combine more than a few factors. This is how he explained why: "With four free parameters one can draw an elephant, with five one can draw an elephant rotating its tail."

IMPERATIVE

When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one.

According to this imperative:
— Do not estimate a density if you need to estimate a function.
(Do not use the classical statistics paradigm for prediction in a high-dimensional world: Do not use generative models for prediction.)
— Do not estimate a function if you only need to estimate its values at given points. *(Try to perform direct inference rather than induction.)*
— Do not estimate predictive values if your goal is to act well.
(A good strategy of action does not necessary rely on good predictive ability.)

3.4.3 RESTRICTIONS ON THE FREEDOM OF CHOICE IN INFERENCE MODELS

In this Afterword we have discussed three levels of restrictions on the freedom of choice in the inference problem:

(1) *Regularization*, which controls the smoothness properties of the admissible set of functions (it forbids choosing an approximation to the desired function from not a "not smooth enough set of functions").

(2) *Structural risk minimization*, which controls the diversity of the set of admissible functions (it forbids choosing an approximation to the desired function from too diverse a set of functions, that is, from the set of functions which can be falsified only using a large number of examples).

(3) *Imperatives*, which control the goals of possible inferences in order to consider a better-posed problem. In our case it means creating the concept of equivalence classes of functions and making an inference using a large equivalence class (it forbids an inference obtained using a "small" equivalence class).

It should be noted that an understanding of the role of a general theory as an instrument to restrict directions of inference has existed in philosophy for a long time. However, the specific formulations of the restrictions as described above were developed only recently. The idea of using regularization to solve ill-posed problems was introduced in the mid-1960s [21, 55]. Structural risk minimization was introduced in the early 1970s [*EDBED*], and the imperative was introduced in the mid-1990s [139].

In order to develop the philosophy of science for a complex world it is important to consider different forms of restriction on the freedom of choice in inference problems and then analyze their roles in obtaining accurate predictive rules for the pattern recognition problem.

One of the main goals of research in the methodology of analysis of a complex world is to introduce new imperatives and for each of them establish interpretations in the corresponding branches of science.

3.4.4 Metaphors for Simple and Complex Worlds

I would like to finish this part of the Afterword with metaphors that stress the difference in the philosophy for simple and complex worlds. As such metaphors let me again use quotes from Albert Einstein.

Two Metaphors for a Simple World

1. *I want to know God's thoughts.* (*A. Einstein*)
2. *When the solution is simple, God is answering.* (*A. Einstein*)

Interpretation

Nature is a realization of the simplest conceivable mathematical ideas. I am convinced that we can discover, by means of purely mathematical constructions, concepts and laws, connect them to each other, which furnish the key to understanding of natural phenomena. (*A. Einstein.*)

Three Metaphors for a Complex World

First Metaphor

Subtle is the Lord, but malicious He is not. (*A. Einstein*)

Interpretation[7]

Subtle is the Lord — one can not understand His thoughts.

But malicious He is not — one can act well without understanding them.

Second Metaphor

The devil imitates God.[8] (*Medieval concept of the devil.*)

Interpretation

Actions based on your understanding of God's thoughts can bring you to catastrophe.

Third Metaphor

If God does exist then many things must be forbidden. (*F. Dostoevsky*)

Interpretation

If a subtle and nonmalicious God exists, then many ways of generalization must be forbidden. The subject of the complex world philosophy of inference is to define corresponding imperatives (to define what should be forbidden). These imperatives are the basis for generalization in real-life high-dimensional problems.

The imperative described in Section 3.4.2 is an example of the general principle that forbids certain ways of generalization.

[7] Surely what Einstein meant is that the laws of nature may be elusive and difficult to discover, but not because the Lord is trying to trick us or defeat our attempts to discover them. Discovering the laws of nature may be difficult, but *it is not impossible*. Einstein considered comprehensibility of the physical world as a "mystery of the world". My interpretation of his metaphor for a *complex world* given below is different.

[8] This includes the claim that for humans the problem of distinguishing imitating ideas of the devil from thoughts of God is ill-posed.

Chapter 4

THE BIG PICTURE

4.1 RETROSPECTIVE OF RECENT HISTORY

The recent history of empirical inference science can be described by Kuhn's model of the development of science which distinguishes between periods with fast development of ideas (development of the new paradigms) and periods with slow developments (incremental research) [168].

In empirical inference science one can clearly see three fast periods: in the 1930s, 1960s, and 1990s.

4.1.1 THE GREAT 1930S: INTRODUCTION OF THE MAIN MODELS

The modern development of empirical inference science started in the early 1930s. Three important theoretical results indicated this beginning:

(1) The foundation of the theory of probability and statistics based on Andrei Kolmogorov's axiomatization and the beginning of the development of the classical theory of statistics.[1]

(2) The development of a basis of applied statistics by Sir Ronald Fisher.[2]

(3) The development of the falsifiability principle of induction by Sir Karl Popper[3].

[1] Andrei Kolmogorov was a leading figure in mathematics in the 20th century. He was the recipient of many of the highest prizes and international awards.

[2] Ronald Fisher was a creator of applied statistics. He was knighted by Queen Elizabeth II in 1952 for his works in statistics and genetics.

[3] Karl Popper was a creator of the modern philosophy of science. He was knighted by Queen Elizabeth II in 1965 for his works in philosophy.

AXIOMATIZATION OF STATISTICS AND PROBABILITY THEORY AND THE PROBLEM OF EMPIRICAL INFERENCE

At the beginning of the 20th century there was a great interest in the philosophy of probabilistic analysis. It was the time of wide discussions about the nature of randomness. These discussions, however, contained a lot of wide speculations. Such speculations were not very useful for development of the formal mathematical theory of random events. In order to separate formal mathematical development of the theory from its interpretation, mathematicians discussed the opportunity to axiomatize the theory of random events. In particular, the problem of the axiomatization of probability theory was mentioned by David Hilbert at the Second Mathematical Congress in Paris in 1900 as one of the important problems of the 20th century.

It took more than 30 years until in 1933 Kolmogorov introduced simple axioms for probability theory and statistics.

Kolmogorov began with a set Ω which is called the *sample space* or set of elementary events A (outcomes of all possible experiments). On the set of possible elementary events a system $\mathcal{F} = \{F\}$ of subsets $F \subset \Omega$, which are called *events*, is defined. He considered that the set $F_0 = \Omega \subset \mathcal{F}$ determines a situation corresponding to an event that always occurs. It is also assumed that the set of events contains the empty set $\emptyset = F_\emptyset \subset \mathcal{F}$, the event that never occurs. Let $\mathcal{F}$ be an algebra of sets. (When $\mathcal{F}$ is also closed under countably infinite intersections and unions, it is called a σ-algebra.) The pair $(\Omega, \mathcal{F})$ defines the *qualitative* aspects of random experiments.

To define the quantitative aspects he introduced the concept called a *probability measure* $P(F)$ defined on the elements F of the set $\mathcal{F}$. The function has the following properties,

$$P(\emptyset) = 0, \quad P(\Omega) = 1, \quad 0 \leq P(F) \leq 1,$$

$$P\left(\cup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i), \quad \text{if } F_i, F_j \in \mathcal{F} \text{ and } F_i \cap F_j = \emptyset \ \ \forall i, j.$$

He then introduced the idea of conditional probability of events

$$P(F|E) = \frac{P(F \cap E)}{P(E)}, \quad P(E) > 0$$

and defined mutual independence of events $F_1, \ldots, F_n$ as the situation when

$$P(F_1 \cap \ldots \cap F_n) = \Pi_{i=1}^{n} P(F_i).$$

By introducing this axiomatization, Kolmogorov made the theory of probabilities a pure mathematical (deductive) discipline with the following basic problem.

THE BASIC PROBLEM OF PROBABILITY THEORY

> Given the triplet $(\Omega, \mathcal{F}, P)$ and an event B, calculate the probability measure $P(B)$.

The axiomatization led to the definition of statistics as the inverse problem.

THE BASIC PROBLEM OF STATISTICS

Given the pair $(\Omega, \mathcal{F})$ and a finite number of i.i.d. data

$$A_1, \ldots, A_\ell$$

estimate the probability measure $P(F)$ defined on all subsets $F \in \mathcal{F}$.

This inverse problem reflects the inductive idea of inference. The general theoretical analysis of inductive inference started with the particular instance of this problem described in the Glivenco–Cantelli theorem (1933) and later was extended to a general theory for uniform convergence (VC theory) in 1968 and 1971.

FISHER'S APPLIED STATISTICS

At about the same time that theoretical statistics was introduced, Fisher suggested the basics of applied statistics. The key element of his simplification of statistical theory was his suggestion of *the existence of a density function* $p(\xi)$ that defines the randomness (noise) for a problem of interest.

Using the density function Fisher introduced the model of observed data

$$(y_1, x_1), \ldots, (y_\ell, x_\ell) \tag{4.1}$$

as measurements of an unknown function of interest $f(x, \alpha_0)$ that belongs to some parametric family $f(x, \alpha), \alpha \in \Lambda$ contaminated by uncorrelated noise defined by the known density $p(\xi)$

$$y_i = f(x_i, \alpha_0) + \xi_i, \quad E\xi_i x_i = 0. \tag{4.2}$$

He developed the maximum likelihood method for estimating the density function given the data (4.1) and the parametric family

$$p_\alpha(\xi) = p(y - f(x, \alpha))$$

that contains the density function of interest. Fisher suggested choosing the density function with parameter α that maximizes the functional

$$R(\alpha) = \sum_{i=1}^{\ell} \ln p(y_i - f(x_i, \alpha)).$$

It took 20 years before LeCam, *using uniform convergence arguments*, proved in 1953 the consistency of the maximum likelihood method for specific sets of parametric families.

Since this time many efforts were made to generalize Fisher's scheme for a wide set of densities (to remove Fisher's requirement to explicitly define the model of noise). In particular, Huber suggested the model of robust estimation that is based on a wide class of density functions. Later, nonparametric techniques also generalized Fisher's model for wide sets of admissible functions.

However, the key element of applied statistics remained the *philosophical realism* based on the generative model (4.2) of the observed data (4.1).

POPPER'S CONCEPT OF FALSIFIABILITY

In the early 1930s Popper suggested his idea of falsifiability. It was considered both as the demarcation line between metaphysics and natural science as well as a justification of Occam's Razor principle. Popper developed the falsifiability idea over his entire lifetime: his first publication appeared in German in 1934 his last addition was in an English edition in 1972. This idea is considered as one of the most important achievements in the philosophy of science of the 20th century.

Fisher's philosophy of applied statistics and Popper's justification of the dependence of generalization ability on the number of entities formed the classical paradigm of philosophical realism for induction.

The continuation of the Kolmogorov–Glivenco–Cantelli line of theoretical statistics led to the development of the VC theory that reflected the philosophical instrumentalism paradigm.

4.1.2 THE GREAT 1960S: INTRODUCTION OF THE NEW CONCEPTS

In the 1960s several revolutionary ideas for empirical inference science were introduced. In particular

1. Tikhonov, Ivanov, and Phillips developed the main elements of the theory of ill-posed problems.
2. Kolmogorov and Tikhomirov introduced the capacity concepts (ε-entropy, covering numbers, width) for sets of functions.
3. Solomonoff, Kolmogorov, and Chaitin developed the concept of algorithmic complexity and used it to justify inductive inference.
4. Vapnik and Chervonenkis developed the basics of empirical inference science.
5. The empirical inference problem became a problem of natural science.

ILL-POSED PROBLEMS

I consider the philosophy of ill-posed problems as the turning point in the understanding of inference. It can have the following interpretation:

(1) *The general problem of inference — obtaining the unknown reasons from the known consequences — is ill-posed.*

(2) *To solve it one has to use very rich prior information about the desired solution. However, even if one has this information it is impossible to guarantee that the number of observations that one has is enough to obtain a reasonable approximation to the solution.*

Therefore one should try to avoid solving an ill-posed problem if possible. The development of the VC theory is just an illustration of this thesis.

THE BASIS FOR AN ALTERNATIVE

The 1960s marked the beginning of the mathematical development of the instrumentalist point of view. In the late 1950s and early 1960s Kolmogorov and Tikhomirov introduced the idea of the capacity of a set of functions and demonstrated its usefulness in function approximation theory.

The VC entropy, Growth function, and VC dimension concepts of capacity (which are different from the one suggested by Kolmogorov and Tikhomirov) became the main concepts that define the generalization ability in the instrumentalist framework.

In the 1960s Solomonoff, Kolmogorov, and Chaitin introduced the concept of algorithmic complexity. Solomonoff introduced this concept in order to understand the inductive principle, while Kolmogorov tried to address issues about nature of randomness,[4] which was a subject of discussion at the beginning of the 20th century.

The book [139] shows that for the pattern recognition problem the idea of Kolmogorov complexity leads to essentially the same bound for the pattern recognition problem that the VC theory gives. The VC theory, however, defines the necessary and sufficient conditions for consistency. It is unclear if algorithmic complexity also provides the necessary and sufficient conditions.

By the end of 1960s we constructed the theory for the uniform law of large numbers and connected it to the pattern recognition problem. By doing this, we had developed the mathematical foundation of predictive learning.

An extremely important fact is that by the end of the 1960s the methodology of solving the inference problem adopted the methodology of the natural sciences. Any results on the generalization problem must be confirmed by computer experiments on a variety of problems.

This forever changed the approach to both empirical inference science and to the philosophy of inference.

4.1.3 THE GREAT 1990S: INTRODUCTION OF THE NEW TECHNOLOGY

In the 1990s the following events took place.

(1) Vapnik and Chervonenkis proved that the existence of the uniform law of large numbers is the necessary and sufficient condition for consistency of the empirical risk minimization principle. This means that if one chooses the function from an admissible set of functions one can not avoid VC type arguments.

(2) Estimation of high-dimensional functions became a practical problem.

(3) Large-margin methods based on the VC theory of generalization (SVM, boosting, neural networks) proved advantageous over classical statistical methods.

These results have led to the development of new learning technologies.

[4] In 1933 when Kolmogorov introduced his axiomatization of probability theory, he effectively stopped these discussions. Thirty years later he came back to this question connecting randomness with algorithmic complexity: Random events are ones that have high algorithmic complexity.

4.1.4 The Great 2000s: Connection to the Philosophy of Science

In the early 2000s the following important developments took place. These all contributed to a philosophy of science for a complex world.

(1) The development of the theory of empirical inference based on VC falsifiability as opposed to Occam's Razor principle.

(2) The development of noninductive methods of inference.

This can lead to a reconsideration of the psychological and behavioral sciences based on noninductive inference. Also it will lead to reconsideration of the goals and methods of pedagogical science: Teaching not just inductive inference but also direct inference that can use the (cultural) Universum.

4.1.5 Philosophical Retrospective

This is how the philosophy of inference has developed.

- At the end of 1930s
  the basics of two different paradigms of empirical inference (generative and predictive) were introduced.

- At the end of the 1960s it became clear that
  the classical statistical paradigm is too restrictive:
  It cannot be applied to high-dimensional problems.

  – At the end of the 1990s it became clear that
  the Occam's Razor principle of induction is too restrictive:
  Experiments with SVMs, boosting, and neural nets provided counter-examples.

- In the beginning of the 2000s it became clear that
  the classical model of science is too restrictive:
  It does not include noninductive (transductive and ad hoc) types of inference which, in high-dimensional situations, can be more accurate than inductive inference.

- In the beginning of the 2000s it also became clear that
  in creating a philosophy of science for a complex world the machine learning problem will play the same role that physics played in creating the philosophy of science for a simple world.

4.2 Large Scale Retrospective

Since ancient times, philosophy has distinguished among three branches: natural science, metaphysics, and mathematics. Over many centuries, there have been ongoing

discussions about the demarcation among these branches, and this has changed many times. Even the existence of these three categories is still under discussion.

Some scholars consider only two categories, including mathematics in the category of natural science or in the category of metaphysics. However, it is convenient for our discussion to consider three categories.

4.2.1 Natural Science

The goal of natural science is to understand and describe the real world. It can be characterized by two features:

> The subject of analysis is defined by the real World.
>
> The methodology is to construct a theory (model) based on the results of both passive (observation-based), and active (query-based) experiments in the real world.

Examples of natural sciences include astronomy, biology, physics, and chemistry where scholars can observe facts of the physical world (conduct passive experiments), ask particular questions about the real world (perform active experiments), and (based on analyzing results of these experiments) create models of the world. In these activities, experiments play a crucial role. From experiments scientists obtain facts that reflect the relationships existing in the world; also, experiments are used to verify the correctness of the theory (models) that are suggested as a result of analysis[5].

Because of this, sometimes natural science is called empirical science. This stresses that the subject of natural science is the real world and its method is the analysis of results of (passive and active) experiments.

4.2.2 Metaphysics

In contrast to empirical (natural) science, metaphysics does not require analysis of experimental facts, or the verification of results of inference. It tries to develop a general way of reasoning with which truth can be found for any imaginary models. Metaphysics stresses the power of pure reasoning.

Examples of metaphysical problems are the following:

> What is the essence of the devil? Here the devil is not necessarily a personalized concept. It can be a metaphor. The metaphors on a complex world used in Section 3.4.4: "The devil imitates God" is one of the concepts of the devil given in the Middle Ages. This was formulated following very wide discussions.
>
> Another example: What is freedom of will?

[5]The relationship between the number of active and passive experiments differs in different sciences. For example, in astronomy there are more passive observations and fewer active experiments. However, in chemistry there are more active experiments and fewer passive experiments.

According to Kant the question

What are the principles of induction?

defines the demarcation line between empirical science (that can be applied only for a particular world) and metaphysics (that can be applied to any possible world). It is commonly accepted that Popper gave the answer to Kant's question by introducing the falsifiability idea. That is, empirical science must be falsifiable.

4.2.3 MATHEMATICS

Mathematics contains elements of both natural science and metaphysics.
There is the following (not quite) jocular definition of pure mathematics (Eugene Wigner):

Mathematics is the science of skillful operation with concepts and rules invented just for this purpose.

From this view, pure mathematics (rather than applied) is a part of metaphysics since mathematics invents concepts and rules for analysis, constructs objects of analysis, and analyzes these objects. It does not rely on experiments either to construct theories or to verify their correctness.

The ideal scheme of pure mathematics is the system that has been used since Euclid introduced his geometry: Define a system of definitions and axioms, and from these deduce a theory.[6] Some scholars, however, consider mathematics as a part of the natural sciences because (according to these scholars) systems of definitions and axioms (concepts and rules) used in mathematics are inspired by the real world.

Another view makes a bridge among mathematics, natural science, and metaphysics by declaring that:

> Everything which is a law in the real world has a description in mathematical terms and everything which is true in mathematics has a manifestation in the real world.

Many scholars consider mathematics as a language that one uses to describe the laws of nature.

We, however, will not discuss the demarcation lines between mathematics and metaphysics, or mathematics and natural science and will instead consider all three as different branches of knowledge.

The duality of the position of mathematics with respect to natural science and metaphysics has important consequences in the history of the development of natural science.

[6]The real picture is much more complicated. According to Israel Gelfand one can distinguish among three periods of mathematical development in the 20th century:

(1) Axiomatization (constructing axioms for different branches of mathematics)

(2) Structurization (finding similar structures in different branches of mathematics)

(3) Renaissance (discovering new facts in different branches of mathematics).

4.3 SHOULDERS OF GIANTS

4.3.1 THREE ELEMENTS OF SCIENTIFIC THEORY

According to Kant, any scientific theory contains three elements:

(1) The setting of the problem,

(2) The resolution of the problem, and

(3) Proofs.

At first glance this remark seems obvious. However, it has a deep meaning. The crux of this remark is the idea that these three elements of the theory in some sense are independent and equally important.

(1) The precise setting of the problem provides a general point of view of the problem and its relation to other problems.

(2) The resolution of the problem comes not from deep theoretical analysis of the setting of the problem but rather precedes this analysis.

(3) Proofs are constructed not to search for the resolution of the problem, but to justify the solution that has already been suggested.

The first two elements of the theory reflect the understanding of the essence of the problem, its philosophy. The proofs make a general (philosophical) model a scientific theory. Mathematics mostly deals with one of these three elements, namely proofs, and much less with setting and resolution.

One interpretation of the Einstein remark:

> *Do not worry about your problems with mathematics, I assure you mine are far greater.*

could be the following:

> The solution of a problem in natural science contains three elements. Proofs are just one of them. There are two other elements: the setting of the problem, and its resolution, which make basis for a theory.

For the empirical inference problem, these three elements are clearly defined:

SETTING.

> The setting of the Inference problems is based on the risk minimization model: Minimize the risk functional
>
> $$R(\alpha) = \int Q(y, f(x, \alpha))dP(y, x), \quad \alpha \in \Lambda$$

in the situation when the probability measure is unknown, but i.i.d. data $(y_1, , x_1), \ldots, (y_\ell, x_\ell)$ are given.

It is very difficult to say who suggested this setting for the first time.[7] I learned about this setting from seminars at the Moscow Institute of Control Science in the mid-1960s.

RESOLUTION.

Two different resolutions of solving this problem were suggested:

(1) Aizerman, Braverman, Rozonoer, and Tsypkin from Russia and Amari from Japan suggested using methods based on gradient-type procedures

$$\alpha_n = \alpha_{n-1} + \gamma_n \text{grad}_\alpha Q(y_n, f(x_n, \alpha_{n-1})).$$

(2) Vapnik and Chervonenkis suggested methods that use the empirical risk minimization principle under the condition of uniform convergence.

PROOFS.

Proofs that justify these resolutions are based on:

(1) The theory of stochastic approximation for gradient based procedures, and

(2) The theory of uniform convergence for the empirical risk minimization principle (1968, 1971). In 1989, Vapnik and Chervonenkis proved that (one-sided) uniform convergence defines the necessary condition for consistency not only for the empirical risk minimization method, but for any method that selects one function from a given set of admissible functions.

4.3.2 BETWEEN TRIVIAL AND INACCESSIBLE

According to Kolmogorov, in the space of problems suggested by the real world there is a huge subspace where one can find trivial solutions. There is also a huge subspace where solutions are inaccessible. Between these two subspaces there is a tiny subspace where one can find non-trivial solutions. Mathematics operates inside this subspace.

It is therefore a big achievement when one can suggest a problem setting and a resolution to this setting and also invent concepts and rules that make proofs both nontrivial and accessible (this is interesting for mathematicians).

In order to transform a problem from an inaccessible one to one that has a mathematical solution very often one must simplify the setting of the problem, perform mathematical analysis, and then apply the result of this analysis to the nonsimplified real-life problem.[8]

[7] I believe that it was formulated by Tsypkin.

[8] In our discussions, the main simplification that made the analysis of induction possible was i.i.d. data in the training and test sets.

Einstein's remark:

> As far as the laws of Mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.

describes this situation.

Nevertheless mathematics forms a universal language for describing the laws of nature that (as we believe) does not contain inner contradictions. There is an understanding that

> The more science uses mathematics, the more truth it contains.

The language, however, is not always equivalent to thought.

4.3.3 Three Types of Answers

Analyzing the real world, mathematics gives three types of answers:

(1) Direct answers,

(2) Comforting answers, and

(3) Tautologies.

(1) A *direct answer* means a direct answer to the posed question. These answers are not necessarily accurate but they are answers to your questions. For example, the answer to the question "How many examples are sufficient to find the ε-approximation to the best possible solution?" is the VC bound. It can be possiby improved, but this is a direct answer to the question.

Few questions have direct answers.

(2) A *comforting answer* does not answer the question of interest since the direct answer is impossible or inaccessible. Instead, it answers another *accessible* question that is somehow related to the question of interest.

For example, one might wonder whether there are enough data to solve a specific ill-posed problem of interest. There is no answer to this question (since it is impossible). Instead mathematics suggests considering a resolution (regularization techniques) for which under some circumstances, in an imaginary (asymptotic) situation, an answer is attainable.

Many more questions have *comforting* answers.

(3) Lastly, mathematics is an instrument that can easily produce many trivial tautologies.[9] As soon as one has a good setting, a decent resolution to this setting, and examples of proofs, one can easily repeat the same construction under slightly different conditions.

It will produce

[9]Many good theorems can be considered as nontrivial tautologies.

formulas, formulas,. . . , formulas

that can be regarded in the same way as Hamlet regarded

words, words,. . . ., words.

Just words, nothing more.

There are many trivial tautologies among the results of mathematical analyses of natural phenomena.

4.3.4 The Two-Thousand-Year-Old War Between Natural Science and Metaphysics

Therefore there is a complicated relationship between metaphysics and natural science which has its reflection in discussions of the role of pure mathematics in natural science.

On the one hand the more mathematics a science uses, the more truth it contains (because we believe that its language does not contain contradictions).

On the other hand, there is a two-thousand-year-old war between metaphysics and natural science.

To discuss the nature of this war, let me start with some well-known quotes that describe it (there are hundreds of similar quotes but these are from intellectual giants):

- I am not a mathematician. I am a natural scientist. *(Kolmogorov, 1973)*
- Theoretical physics is too difficult for physicists. *(Poincare, 1910)*
- A mathematician may say anything he pleases, but a physicist must be at least partially sane. *(Gibbs, 1889)*
- I have hardly ever known a mathematician who was able to reason. *(Plato, 370 BC)*

Why did Kolmogorov not like to call himself a mathematician?[10] Why didn't Plato take mathematicians as seriously as the natural philosophers (people involved in discussion about fundamental principles of nature)?

[10]Kolmogorov did not play with formulas. The concepts and rules that he introduced in different branches of mathematics (probability theory, information theory, theory of approximation, functional analysis, logic, differential equations) helped to advance philosophy in natural science. These are some examples of his ideas related to the subject of this book:

He obtained the bound whose generalization is the bound on the uniform law of large numbers.

He introduced the concept of ε-entropy which provided the opportunity to consider capacity concepts of learning theory and in particular the VC entropy.

His idea of algorithmic complexity was used in the minimum message (minimum description) length principle leading to learning methods with the same generalization bound as the VC bound [139].

Kolmogorov did not work on pattern recognition problems but he developed the concepts and rules that were very similar to the one behind the main philosophy of learning theory.

This could be the answer. Natural science is not only about proofs but more about the setting and resolution of problems. Mathematics is just a language that is useful for the setting, the resolutions, and especially for the proofs. To use this language well requires a high level of professionalism. This is probably what Poincare had in mind when he made the above-quoted remark.[11]

However, to find a good setting and a good resolution requires another sort of professionalism. I believe the tendency to underestimate the role of this sort of professionalism and overestimate the role of technical (mathematical) professionalism in analysis of nature was the reason for Plato's remark.

The research in empirical inference science requires searching for new models of inference (different from inductive inference, such as inference in Universum environment, transductive, selective, ad hoc inferences, and so on). They are currently not under the scope of interest of mathematicians since they do not yet have clear settings and clear resolutions (this is the main subject of research). Mathematicians will become interested in this subject later when new settings, new resolutions, and new ideas for proofs are found.

The goal of the empirical inference discipline is to find these elements of the theory.

4.4 To My Students' Students

4.4.1 Three Components of Success

To be successful in creativity, and in particular in scientific creativity, one has to possess the following three gifts:

(1) Talent and strong motivation,

(2) Ability to work hard, and

(3) Aspiration for perfection and uncompromising honesty to one's inner truth.

Most discussions about the components of success concentrate on the first two gifts. One can easily recognize them observing the work of an individual (how bright the individual is in solving problems, how fast he understands new concepts, how many hours he works, and so on). These two components form the *necessary* conditions for success.

The third (maybe the most important) component that provides the spirit of creativity, the inner quality control for creation, a concept of high standards, and the willingness to pay any price for this high standard is more delicate. It can not be seen as easily as the first two. Nevertheless, it is the demarcation line between individuals whose lifetime achievements are above the expectations of their colleagues and individuals whose lifetime achievements are below the expectations of their colleagues.

In the next section I will try to describe this gift and to show that when creating something new one encounters two problems:

[11] By the way, the main part of theoretical physics was done by theoretical physicists who sometimes used "dirty mathematics." Later, mathematicians cleaned up the mathematics.

(1) to develop one's ideas in the way one desires, and

(2) to develop them perfectly.

The second of these two problems requires the most effort.

4.4.2 The Misleading Legend About Mozart

There is a highest standard of genius: Mozart, the greatest wunderkind, the greatest musician, and the greatest composer of all time. The legend gives the impression that everything Mozart touched achieved perfection automatically, without much effort. In many languages there is an expression "Mozart's lightness."

Legend admits, however, that when he was very young he worked extremely hard (he did not have a normal childhood; he was under very strong pressure from his father, who forced him to practice a lot).

Then legend tells us about Mozart, a merry young man, visiting a variety of Vienna cafes, who had admirers in everyday life, yet created the greatest music. He wrote it down with no draft.

That is true, Mozart did not write any drafts. He possessed a phenomenal professional memory, created his compositions in his mind, and could work simultaneously on several different compositions. Because of this, it would seem that his creativity also came easily. This was not the case. Legend tells us stories that he was almost always late in finishing the masterpieces which he committed to create.

The work which no one saw that he did in his mind was so exhausting that Mozart sometimes was not able to speak; he barked like a dog and behaved inappropriately sometimes like an idiot. He badly needed relaxation from his inner work, therefore he visited Vienna cafes (the simpler, the better) where he looked for a break from his exhausting concentration. He almost killed himself by such work that no one could see. By the end of his life (he was only 35 years old when he died) he was a very sick person who had used up his life: he had no time to properly build his family life (he married, almost by chance, the daughter of his landlord) and he had no time for friendship. He gave up everything for his genius.

One can say that this is just speculation; no one can tell you what was going on inside Mozart. That is true. But fortunately there is a recording made for Deutsche Gramophon called the "Magic of Horowitz." In this album there are 2 CDs and one bonus DVD. The DVD documents the recording of Mozart's 23rd piano concerto, played by the pianist of the century, Vladimir Horowitz. He is accompanied by one of the world's best orchestras, the orchestra of Teatro alla Scallo, under the direction of maestro Carlo Maria Giulini.

I believe that Horowitz's interpretation of Mozart's work and his uncompromised demand for excellence reflects Mozart's spirit for perfection.

4.4.3 Horowitz's Recording of Mozart's Piano Concerto

The record was made in 1987 in Horowitz's final years (he was 84). You see an old man who can hardly walk, and who probably has different health problems but who does not forget for a second about the necessity to play perfectly. In spite of all of his past achievements, he is not sure he will succeed. He asks his assistant (who turns the pages of his score), "Are my fingers good?" When one of the visitors (who came from England to Italy just to see this recording) tells him that she likes his tie he immediately reacts, "Do you like my tie more than my playing?" and repeats this several times throughout the session.

Probably the best part of this DVD is when, after recording the glamorous second movement, Horowitz, Giulini (who was selected by Horowitz for this record), and their producer evaluate the recording. You see the pianist's striving for perfection, the conductor's uncertainty, and how long it takes them to relax and agree that the record is good.

At the start of the third movement, Horowitz and the orchestra did not play together perfectly. The producer immediately stopped recording and suggested repeating it. You see how the great Horowitz without any doubt accepts his part of the fault and then how deeply he concentrated and how wonderfully he performed on the last movement in the next recording attempt.

Then he chats with musicians that came from all over the World just to see this session. The very last words of Horowitz on the DVD were his recollection of excellent reviews on the previous performance. However, he immediately added, "But this makes no difference."

That is, it does not matter how good you were yesterday; it makes no difference for today's results. Today a new challenge starts. That is the way of all great intellectual leaders.

4.4.4 Three Stories

Since ancient times people saw a very specific relationship between a genius and his professional work. Cicero formulated this as follows:

> *Among all features describing genius the most important is inner professional honesty.*

There are a lot of examples where the moral quality of a genius in everyday life does not meet high standards, but they never lose these high standards in their professional lives.

A person who plays games with professional honesty loses his demand for inner truth and compromises with himself. This leads to a decrease in his ability to look for the truth. Let me give examples of actions of my heros Kolmogorov, and Einstein.

Kolmogorov. There is a legend that Kolmogorov read everything. Nobody knows when and how he accomplished it, but somehow he did. In the beginning of the 1960s an unknown young researcher Ray Solomonoff, working for a small private company

in Boston released a report titled, " A Preliminary Report on a General Theory of Inductive Inference" [171]. This report contained ideas on inference and algorithmic complexity. Probably very few researchers read this report but Kolmogorov did. In 1965 Kolmogorov published his seminal paper where he introduced the concept of Kolmogorov complexity to answer the question "What is the nature of randomness?"

In this article he wrote that he was familiar with Solomonoff's work and that Solomonoff was the first to suggest the idea of algorithmic complexity.

In 2002 Solomonoff became the first person awarded the Kolmogorov medal established by the Royal Holloway University of London.

Einstein. In 1924 Einstein received a letter from the Indian researcher Bose which contained the handwritten manuscript "Planck's Law and the Hypothesis of Light Quanta," written in English. Einstein translated this letter into German and submitted it to the *Zeitschrift für Physik* with a strong recommendation. This was the main work of Bose. Later Einstein significantly extended the ideas of Bose and published several papers on this subject.

One can say that this just reflects human decency. Not only that. They also did these things to keep themselves honest in order not to betray their own individuality. The smallest compromise here leads to a compromised demand of yourself, which leads to a decline in creativity. They also did it because they had responsibility before their talents.

Galois. The last story is about Evariste Galois which I have been trying to understand ever since I read about it. This story is about the responsibility of the great talent with respect to results of his work.

Galois was a very talented mathematician and squabbling young man who during almost all of his short life produced nothing but trouble. As a result, this kid entered into a duel and was killed when he was just 20. The night before this duel he wrote down the mathematical theory that is now called "Galois Theory."

Why? He was not a stupid kid. He understood the consequences of not sleeping the night before a duel. Why didn't he think like this: "I must sleep to perform well tomorrow. This, is the most important thing. If I do not die I will write down the theory later, if I do die who cares?"

Why did this kid, who looked like just a troublemaker, accepted such a big responsibility? It was something bigger than himself. As with all great people he had a burden of responsibility that came along with his talent and Galois paid the full price of his life for this. He belonged not only to himself.

He was very different from most people.

4.4.5 Destructive Socialist Values

People are not born equal in their potential (they are not identical). Among those who are born there are future beautiful women, and future gentlemen, future musicians, and future scientists, there are very warm family people (this is also a great talent), and there are misanthropes. Among these who are born are future Mozarts and Einsteins. They

are all very different, and they bring a variety of human talents into society. People are not equal in their abilities.

This inequality of individuals inspired strong negative feelings among people who identify themselves as socialists. Socialist discussions on equality of individuals have continued for thousands of years. The main subject of these discussions has not been how to distribute wealth justly (this is a common misconception) but how to make people equal (essentially identical). One can find reflections of these discussions in philosophy, literature, social movements, and state systems. The practical implications of socialist ideas always were attempts to make people identical *by force*, not allowing them to be too different from an accepted standard.

A deep analysis of this phenomena was done by Igor Shafarevich in his book *The socialist phenomenon* [151].[12] The main results of his analysis is that the impulse to seek utopia (identity of individuals) is deeply rooted in the human psyche. This impulse, however, is very dangerous, because it leads to suppression of individuals in favor of unificated community and this in turn leads to societal decline.

Over thousands of years (ranging from ancient Mesopotamia and medieval Inca Empire to recent Russia and Cambodia) the socialist ideas has always brought humans to catastrophic consequences, but it has nonetheless always resurfaced, despite all past experience. For reasons that are unclear it can be attractive even to high-level intellectuals (Plato supported it in his *Republic*).

But despite the miserable failure of *all* classical socialist systems, its ideas remain attractive for many at the level of *socialist values*. This is how *The Concise Oxford Dictionary of Current English* [172] defines the essence of these values:

> Socialism is a principle that individual freedom should be completely subordinated to interest of community, with any deductions that may be correctly or incorrectly drawn from it.[13]

Renaming the Vapnik–Chervonenkis lemma as the Sauer lemma and the attempt to create the PAC legend described in Section 2.1 were beneficial to no one personally. These were the actions that execute the main slogan of a socialist community:

Expropriate extras and split them equally.

Socialist values have a very negative effect in science since they lead to a strong resistance to original ideas in favor of ones shared by a community.[14]

[12] I believe that one cannot consider his/her education complete without reading this book. It shows that the socialist phenomenon based on the idea of fundamental identity of individuals (udenticalness of individuals) existed from very ancient times and constitutes one of the basic forces of history. It is the instinct of *self-destruction* of a society that has a strong attraction for some of its parts (similar, for example, to a strong attraction for some individuals to jump down being on an edge of a very high place).

[13] I have had experience of the strong pressure of this principle both in Soviet Russia and in the United States. However below I refer only to my limited experience in the United States.

[14] As an illustration of this statement let me quote from three reviews on three different proposals that suggested to develop the ideas described in this Afterword and rejected by the US National Science Foundation (NSF) based on socialist arguments (I emphasize them in bold font).

REVIEW 1. Release date 07/21/2004. (On Science of Learning Center.)
The described intellectual merit is substantial and impressive. This proposal purports to break fundamentally new ground in our approach to scientific reasoning and developing powerful new algorithms. The claimed

This creates (using Popper's words [150, VIII]) *"the fundamental subjectivist position which can conceive knowledge only as a special kind of mental state, or as a disposition, or as a special kind of belief.*[15]

My understanding of the reason why:

> *Great spirits have always encountered violent opposition from mediocre minds*
> (A. Einstein)

is because the great spirits are *unique*, and this contradicts the socialist belief of mediocracy, the fundamental identity of individuals.

---

impact would be significant both in practice and in our fundamental view of science.

This is an extremely well written, argued, and engineered proposal for a center with a strong theme. It is a pleasure to read, and highly convincing. There is a level of excitement and importance that is communicated. This list of practical implications, in itself, is well worth the effort, but the theoretical ones are rather intriguing. These are highly qualified scientists asking for funding to accelerate development toward a new kind of reasoning. One does not routinely read a proposal like this.

Unfortunately the case for "centerness" is weak inasmuch as the work is rather narrowly structured, and poorly argued for in some of the critical characteristics. **The intellectual atmosphere might be enriched for the students if there were a great diversity of viewpoints, and it would enhance the case for forming a center.**

FROM REVIEW 2. Release date 06/23/2005. (On Empirical Inference in Social and Behavioural Science.) Dr. Vapnik proposes to develop a philosophy of science for the complex world using ideas from statistical learning theory, support vector machines, and transductive reasoning. The proposal was actually fascinating to read, and I thought the background information on SLT was especially clear. However, the scope of the proposal is enormous if the aim is to construct such a comprehensive philosophy and relate it to the social science. **One thing that strikes me is that this philosophy of science and ideas related to generalizability would all be centred around Dr. Vapnik's personal theoretical contributions to SLT.**

FROM REVIEW 3. Release date 09/17/2005. (On Directed Ad-Hoc Inference.)
Although the idea of developing a new type of inference is very interesting, **the proposed methodology heavily relies on the work previously done by the PI on support vector machines, and non-parametric estimation of conditional probabilities.**

The projects were rejected not because there were doubts about their scientific significance but because they are based on ideas that do not represent a widespread community.

[15] As an example of what is *to conceive knowledge as a belief* let me quote from another NSF review (on another proposal) that contains no arguments but agressive attack from a position of blind corporate belief (the quote from the review I made in bold font and my comments in parentheses).

FROM REVIEW 4. Release date 01/31/06. (On Relation to the Philosophy of Science.)
The discussion of philosophy of science is breathtakingly naive including an over-simplistic account of simplicity and complexity (this is about advanced complexity concepts, the VC dimension and VC entropy; is there something better? *V.V.*), **a completely implausible characterization of key methodological difference between physics and social science** (this is about Einstein's remark that the methodology of classical science cannot be used when one must consider too many factors and an attempt to create methodology for such situations *V.V.*), **and promise to give that dead-horse Popperian concept of corroboration another whipping.** (Sir Karl Popper was the first who tried to justify induction using the idea of capacity of set (falsifiability). He made mistakes. However, another capacity concept (the VC falsfiability) does define the necessary and sufficient conditions for predictive induction. *V.V.*). **Moreover, there is virtually no reference to pertinent philosophical discussion such as . . .** (two irrelevant traditional works *V.V.*).

It is sad to see the same collective socialist logic in all above reviews: *ideas of individuals should be completely subordinated to interest of community, with any deductions that may be correctly or incorrectly drawn from it.*

This is why there exist rude attack against great spirits demonstrating disrespect to them and in particular against Einstein as the brightest figure in scientific originality ("Many of his ideas were suggested by his wife." "He was not original since similar ideas were suggested by Poincare, Lorenz, and Minkowski." "He was a terrible family man," and so on). The same sort of criticism exists even against Sir Isaac Newton ("He spent most of his life working on stupid things and was a terrible man"). The main message is:

" Look at them, they are almost no different from us."

Maybe it is true that sometimes they behave like us (or maybe even worse than us). But they are very different in their vision of the truth, their devotion to this truth, and their honesty in pursuing the truth.

4.4.6 Theoretical Science Is Not Only a Profession — It Is a Way of Life

Being a natural science theorist is not only a profession, but it is also a difficult way of life:

— You come into this world with your individual seed of truth.

— You work hard to make your truth clear.

— You push yourself to be unconditionally honest with respect to your truth.

— You have a life-long fight protecting your truth from old paradigms.

— You resist strong pressure to betray your truth and become part of a socialist community.

If you have a talent, the character to bear such a life, and a little luck, then you have a chance to succeed: To come into this world bringing your own seed of truth, to work hard to make your truth clear, and to add it to the Grand Truth.

There is a warm feeling of deep satisfaction for those who have made it. And even if one cannot call this genuine happiness, it can be a very good substitute.

BIBLIOGRAPHY

For items with numbers 1 – 120 see pages 391 – 396.

121 Vapnik, V., and Chervonenkis, A.: The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Yearbook of the Academia of Sciences of USSR on Recognition, Classification, Forecasting* Vol 2, Moscow, Nauka, 207–249 (in Russian), 1989.

English translation: Vapnik, V., and Chervonenkis, A.:The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recogn. Image Anal.* **1** (3), 284-305, 1991.

122 Valiant, L.: Theory of the learnability. *Commun. ACM* **27**, 1134–1142, 1984.

123 Valiant,L.: A view of computational learning theory. In *Computation & Cognition*, Proceeding of the First NEC Research Symposium, Society for Industrial and Applied Mathematics, Philadelphia, 32–51, 1990.

124 Gurvits, L.: A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. In Lee and Maruoka (Eds), *Algorithmic Learning Theory* ALT-97, LNAI-1316, Berlin, Springer, 352–363, 1997.

125 Boser, B., Guyon, I., and Vapnik, V.: A training algorithm for an optimal margin classifier. *Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh AMC, 144–152, 1992.

126 Aizerman, M., Braverman, I., and Rosonoer, L.: Theoretical foundations of the potential functions method in pattern recognition learning. *Automation and Remote Control* **25**, 821–837.

127 Wallace, C., and Boulton, D.: An information measure for classification. *Comput. J.* **11**, 85–95, 1968.

128 Rissanen, J.: Modelling by shortest data description. *Automatica* **14**, 465–471, 1978.

129 Kolmogorov, A.: Three approaches to the quantitative definition of information. *Prob. Inf. Transm.* **1** (1), 1–7, 1965.

130 Sauer, N.: On the density of families of sets. *J. Comb. Theor. (A)* **13**, 145–147, 1972.

131 Shelah, S.: A computational problem: Stability and order of models and theory of infinitary languages. *Pacific J. Math.* **41**, 247–261, 1972.

132 Cortes, C., and Vapnik, V.: Support vector networks. *Mach. Learn.* **20**, 1–25, 1995.

133 Lugosi, G., and Zeger, K.: Concept of learning using complexity regularization. *IEEE Trans. on Information Theory* **41**, 677–678, 1994.

134 Devroye, L., and Wagner, T.: Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics* **8**, 231–239, 1982.

135 Freund, Y. and Schapire, R.: Experiments with new boosting algorithms. In *Proceedings of the 13th International Conference on Machine Learning*, San Francisco: Morgan-Kaufmann, 148–156, 1996.

136 Schapire, R., Freund, Y., Bartlett, P., and Lee, W.: Boosting the margin: A new explanation for effectiveness of voting methods. *Ann. Stat.* **26**, 1651–1686, 1998.

137 Popper, K.: *The Logic of scientific discovery*(2nd ed.). New York: Harper Torch, 1968.

138 Shawe-Taylor, J., Bartlett, P., Williamson, C., and Anthony, M.: Structural risk minimization over data-dependent hierarchies. IEEE Trans. Info. Theor. **44** (5), 1926–1940, 1998.

139 Vapnik, V.: *The nature of statistical learning theory*. New York: Springer, 1995.

140 Vapnik, V.: *Statistical learning theory*. New York: Wiley, 1998.

141 Alon, N., Ben-David, S., Cesa-Bianch, N., and Haussler, D.: Scale-sensitive dimension, uniform convergence, and learnability. *J. ACM* **44**(4), 615–631, 1997.

142 Bottou, L., Cortes, C., and Vapnik, V.: On the effective VC dimension. Tech. Rep. Neuroprose, (ftp://archive.cis.ohio–state.edu/pub/neuroprose), 1994.

143 Vapnik, V., and Chervonenkis, A.: Uniform convergence of frequencies of occurrence of events to their probabilities. *Dokl. Akad. Nauk USSR* **181**, 915–918, 1968. (English translation: *Soviet Math. Dokl.* **9**, 4, (1968).)

144 Vapnik, V., Levin, E., and LeCun, Y.: Measuring the VC dimension of a learning machine. *Neural Computation* **10** (5), 1994.

145 Bottou, L., and Vapnik, V.: Local learning algorithms. *Neural Computation* **6** (6), 888–901, 1992.

146 Weston, J., Perez-Cruz, F., Bousquet, O., Chapelle, O., Elisseeff, A., and Shölkopf, B.: KDD cup 2001 data analysis: prediction of molecular bioactivity for drug design—binding to thrombin. *Bioinformatics*, 2003.

147 Nadaraya, E.: On estimating regression. *Theor. Prob. Appl.* **9**, 1964.

148 Watson,G.: Smooth regression analysis, *Shankhaya*, Seria A **26**, 1964.

149 Stute, W.: On almost sure convergence of conditional empirical distribution function, *Ann. Probab.* **14**(3), 891–901, 1986.

150 Popper, K.: *Conjectures and Refutations*. New York: Routledge, 2000.

151 Shafarevich, I.: *The Socialist phenomenon*. New York: Harper and Row, 1980. (Currently this book is out of print. One can find it at http://www.robertlstephens.com/essays/shafarevich/001SocialistPhenomenon.html.)

152 Schölkopf, B., and Smola, A.: *Learning with kernels*, Cambridge MA: MIT Press, 2002.

153 Cristianini, N., and Shawe-Taylor, J.: *An introduction to Support Vector Machines*. Cambridge: Cambridge University Press, 2000.

154 Joachims, T.: *Learning to classify text using support vector machines*, Hingham, MA: Kluwer Academic Publishers, 2002.

155 Herbrich, R.: *Learning kernel classifiers: theory and algorithms*. Cambridge MA: MIT Press, 2002.

156 Shawe-Taylor, J., and Cristianini, N.: *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press, 2004.

157 Abe, S.: *Support vector machines for pattern classification*. New York: Springer, 2005

158 Schölkopf, B., Burges, C., and Smola, A. (Eds.): *Advances in Kernel Methods. Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

159 Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D. (Eds.) *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 2000.

160 Schölkopf, B., Tsuda, K., and Vert, J. (Eds.): Kernel Methods in Computational Biology. Cambridge, MA: MIT Press, 2004.

161 Chapelle, O., Schölkopf, B., and Zien, A. (Eds.): *Semi-supervised Learning*, Cambridge MA: MIT Press, 2006.

162 Wang, L. (Ed.): *Support vector machines: theory and applications*, New York: Springer, 2005.

163 Lee, S.-W, and Verri, A.(Eds.): *Pattern recognition with support vector machines*, New York: Springer, 2002.

164 Schölkopf, B., and Warmuth, M. (Eds.): *Learning theory and kernel machines.* New York: Springer, 2003.

165 Burges, C.: A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, **2**(2), 121–167, 1995.

166 Phillips, D.: A technique for numerical solution of certain integral equation of first kind. *J. Assoc. Comput. Math.* **9**, 84–96, 1962.

167 Steinwart, I.: Consistency of Support Vector Machines and other regularized kernel machines. *IEEE Trans. Info. Theor.* , **51**, 128–142, 2005.

168 Kuhn, T.: *The Structure of Scientific Revolutions* (3rd ed.). Chicago: Chicago University Press, 1996.

169 Hempel, K.: *The philosophy of K. Hempel: Studies in Science, Explanation, and Rationality*. Oxford, UK: Oxford University Press, 2001.

170 Chaitin, G.: On the length of programs for computing finite binary sequences. *J. Assoc. Comput. Math.* **4**, 547–569, 1996.

171 Solomonov, R.: A preliminary report on general theory of inductive inference, Technical Report ZTB–138, Cambridge, MA: Zator Company, 1960.

172 Fowler, H.W. (Ed.): The Concise Oxford Dictionary of Current English, Oxford, UK: Oxford University Press, 1956.

173 Wapnik, W.N., Tscherwonenkis, A. Ya: *Theorie der Zeichenerkennung*. Berlin: Akademia, 1979. (German translation of the book [12]).

174 Chapelle O. and Zien A.: Semi-supervised classification of low density separation. Proc. of the Thenth International Workshop on Artificial Intelligence and Statistics, pp 57 – 64, 2005

INDEX