1

TWO APPROACHES TO THE LEARNING PROBLEM

In this chapter we consider two approaches to the learning problem—the problem of choosing the desired dependence on the basis of empirical data.

The first approach is based on the idea that the quality of the chosen function can be evaluated by a risk functional. In this case the choice of the approximating function from a given set of functions is a problem of minimizing the risk functional on the basis of empirical data. This problem is rather general. It embeds many problems of statistics. In this book we consider three of them: pattern recognition, regression estimation, and density estimation.

The second approach to the learning problem is based on estimating desired stochastic dependencies (densities, conditional densities, conditional probabilities). It requires solution of integral equations (determining these dependencies) in situations where some elements of the equations are known only approximately. Using estimated stochastic dependence, the pattern recognition and regression estimation problems can be solved as well. However, the function obtained by solution of the integral equations provides much more details than is required for these problems. The price we pay for these details is the necessity to solve ill-posed problems.

1.1 GENERAL MODEL OF LEARNING FROM EXAMPLES

Consider the following model of searching for functional dependency, which we call the *model of learning from examples*.

The model contains three elements (Fig 1.1):

1. The generator of the data (examples), G.

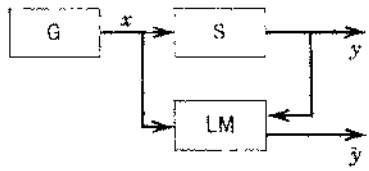

FIGURE 1.1. A model of learning from examples. During the learning process, the learning machine observes the pairs (x,y) (the training set). After training, the machine must on any given x return a value $\hat{y}$. The goal is to return a value $\hat{y}$ which is close to the supervisor's response y.

2. The target operator (sometimes called *supervisor's operator* or, for simplicity, *supervisor*), S.
3. The learning machine, LM.

The generator G is a source of situations that determines the environment in which the supervisor and the learning machine act. In this book, we consider the simplest environment: G generates the vectors $x \in X$ *independently and identically distributed* (i.i.d.) according to some unknown (but fixed) probability distribution function $F(x)$.

These vectors are inputs to the target operator (supervisor); the target operator returns the output values y. The target operator, which transforms the vectors x into values y, is unknown, but we know that it exists and does not change.

The learning machine observes ℓ pairs

$$(x_1, y_1), \ldots, (x_\ell, y_\ell)$$

(the *training set*) which contain input vectors x and the supervisor's response y. During this period, the learning machine constructs some operator which will be used for prediction of the supervisor's answer y_i on any specific vector x_i generated by the generator G. The goal of the learning machine is to construct an appropriate approximation.

To be a mathematical statement, this general scheme of learning from examples needs some clarification. First of all, we have to describe what kind of operators are used by the supervisor. In this book, we suppose that the supervisor returns the output y on the vector x according to a conditional distribution function $F(y|x)$ (this includes the case when the supervisor uses some function $y = f(x)$).

Thus, the learning machine observes the training set, which is drawn randomly and independently according to a joint distribution function $F(x,y) = F(x)F(y|x)$. (Recall that we do not know this function but we do know that it exists.) Using this training set, the learning machine constructs an approximation to the unknown operator.

To construct an approximation, the learning machine chooses one of the

two goals to pursue:

- To *imitate* the supervisor's operator: Try to construct an operator which provides for a given generator G, the best prediction to the supervisor's outputs.
- To *identify* the supervisor's operator: Try to construct an operator which is close to the supervisor's operator.

There exists an essential difference in these two goals. In the first case, the goal is to achieve the best results in prediction of the supervisor's outputs for the environment given by the generator G. In the second case, to get good results in prediction is not enough; it is required to construct an operator which is close to the supervisor's one in a given metric. These two goals of the learning machine imply two different approaches to the learning problem.

In this book we consider both approaches. We show that the problem of imitation of the target operator is easier to solve. For this problem, a nonasymptotic theory will be developed. The problem of identification is more difficult. It refers to the so-called ill-posed problems. For these problems, only an asymptotic theory can be developed. Nevertheless, we show that the solutions for both problems are based on the same general principles.

Before proceeding with the formal discussion of the learning problem, we have to make a remark. We have to explain what it means "to construct an operator" during the learning process. From a formal point of view, this means that the learning machine can implement some fixed set of functions given by the construction of the machine. During the learning process, it chooses from this set an appropriate function. The rule for choosing the function is one of the most important subjects of the theory and it will be discussed in this book. But the general assertion is:

The learning process is a process of choosing an appropriate function from a given set of functions.

We start our discussion of the learning problem with the problem of imitation. It is based on the general statistical problem of minimizing the risk functional on the basis of empirical data. In the next section we consider a statement of this problem, and then in the following sections we demonstrate that different learning problems are particular cases of this general one.

1.2 THE PROBLEM OF MINIMIZING THE RISK FUNCTIONAL FROM EMPIRICAL DATA

Each time the problem of selecting a function with desired quality arises, the same model may be considered: Among the totality of possible functions, one

looks for the one that satisfies the given quality criterion in the best possible manner.

Formally this means that on the subset Z of the vector space R^n, a set of admissible functions $\{g(z)\}, z \in Z$, is given, and a functional

$$R = R(g(z)) \tag{1.1}$$

is defined which is the criterion of quality of the chosen function. It is then required to find the function $g^*(z)$ from the set $\{g(z)\}$ which minimizes 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 set of functions $\{g(z)\}$ and the functional $R(g(z))$ are explicitly given, the search for the function $g^*(z)$ which minimizes $R(g(z))$ is the subject of the calculus of variations.

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

$$R(g(z)) = \int L(z, g(z))\, dF(z), \tag{1.2}$$

where function $L(z, g(z))$ is integrable for any $g(z) \in \{g(z)\}$. The problem is to minimize the functional (1.2) in the case when the probability distribution $F(z)$ is unknown but the sample

$$z_1, \ldots, z_\ell \tag{1.3}$$

of observations drawn randomly and independently according to $F(z)$ is available.

Sections 1.3, 1.4, and 1.5 shall verify that the basic statistical problems related to function estimation problem can be reduced to the 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 directly 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 $g^*(z)$ from the set $\{g(z)\}$ which minimizes (1.1). When (1.2) is to be minimized on the basis of empirical data (1.3), the basic problem is to formulate a constructive criterion for choosing the function rather than organizing the search of the functions in $\{g(z)\}$. (The functional (1.2) by itself cannot serve as a selection criterion, since the measure $F(z)$ involved in it is unknown.) Thus, in the first case, the question is:

How can we obtain the minimum of the functional in the given set of functions?

While in the second case the question is:

What should be minimized in order to select from the set $\{g(z)\}$ a function which will guarantee that the functional (1.2) is small?

Strictly speaking, one cannot minimize (1.2) based on (1.3) using methods developed in optimization theory. The minimization of the functional (1.2) on the basis of empirical data (1.3) is one of the main problems of mathematical statistics.

When formulating the minimization problem for functional (1.2), the set of functions $g(z)$ will be given in a parametric form $\{g(z,\alpha),\ \alpha \in \Lambda\}$.† Here α is a parameter from the set Λ such that the value $\alpha = \alpha^*$ defines the specific function $g(z,\alpha^*)$ in the set $g(z,\alpha)$. Finding the required function means determining the corresponding value of the parameter $\alpha \in \Lambda$.

The study of only parametric sets of functions is not a restriction on the problem, since the set Λ, to which the parameter α belongs, is arbitrary: It can be a set of scalar quantities, a set of vectors, or a set of abstract elements.

In the new notation the functional (1.2) can be rewritten as

$$R(\alpha) = \int Q(z,\alpha)\, dF(z), \qquad \alpha \in \Lambda, \tag{1.4}$$

where

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

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

The problem of minimizing functional (1.4) 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 the vector z. The *expected loss* (with respect to z) for the function $Q(z,\alpha^*)$ is determined by the integral

$$R(\alpha^*) = \int Q(z,\alpha^*)\, dF(z).$$

This functional is called the *risk functional* or the *risk.* The problem is to choose in the set $Q(z,\alpha)$, $\alpha \in \Lambda$, a function $Q(z,\alpha_0)$ which minimizes the risk when the probability distribution function is unknown but random independent observations $z_1, \ldots, z_\ell$ are given.

Remark. Let us clarify the phrase "probability distribution function is unknown." Denote by $\mathcal{P}_0$ the set of all possible probability distribution functions on Z and by $\mathcal{P}$ some subset of probability distribution functions from $\mathcal{P}_0$.

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

We will distinguish between two cases:

1. Case where we have no information about the unknown distribution function. (We have only the trivial information that $F(z) \in \mathcal{P}_0$.)
2. Case where we have nontrivial information about the unknown distribution function. We know that $F(z)$ belongs to the subset $\mathcal{P}$ which does not coincide with $\mathcal{P}_0$.

In this book, we consider mostly the first case, where we have no a priori information about the unknown distribution function. However, we will consider the general method for constructing a theory which is valid for any given set of probability measures.

The problem of minimizing the risk functional (1.4) on the basis of empirical data (1.3) is rather general. It includes in particular three basic statistical problems:

1. The problem of pattern recognition
2. The problem of regression estimation
3. The problem of density estimation

In the next sections we shall verify that all these problems can be reduced to the minimization of the risk functional (1.4) on the basis of the empirical data (1.3).

1.3 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 supervisor observes occurring situations and determines to which of k classes each one of them belongs. It is required to construct a machine which, after observing the supervisor's classification, carries out the classification approximately in the same manner as the supervisor.

Using formal language, this statement can be expressed as follows: In a certain environment characterized by a probability distribution function $F(x)$, situation x appears randomly and independently. The supervisor classifies each situations into one of k classes. We assume that the supervisor carries out this classification using the conditional probability distribution function $F(\omega|x)$, where $\omega \in \{0, 1, ..., k-1\}$ ($\omega = p$ indicates that the supervisor assigns situation x to the class number p).†

† This is the most general case which includes a case when a supervisor classifies situations x using a function $\omega = f(x)$.

Neither the properties of the environment $F(x)$ nor the decision rule of the supervisor $F(\omega|x)$ are known. However, we do know that both functions exist. Thus, a joint distribution $F(\omega, x) = F(\omega|x)F(x)$ exists.

Now, let a set of functions $\phi(x, \alpha)$, $\alpha \in \Lambda$, which take only k values $\{0, 1, ..., k-1\}$ (a set of decision rules), be given. We shall consider the simplest loss function

$$L(\omega, \phi) = \begin{cases} 0 & \text{if } \omega = \phi \\ 1 & \text{if } \omega \neq \phi. \end{cases} \tag{1.5}$$

The problem of pattern recognition is to minimize the functional

$$R(\alpha) = \int L(\omega, \phi(x, \alpha))\, dF(\omega, x) \tag{1.6}$$

on the set of functions $\phi(x, \alpha)$, $\alpha \in \Lambda$, where the distribution function $F(\omega, x)$ is unknown but a random independent sample of pairs

$$(\omega_1, x_1), ..., (\omega_\ell, x_\ell) \tag{1.7}$$

is given. For the loss function (1.5), the functional (1.6) determines the probability of a classification error for any given decision rule $\phi(x, \alpha)$.

The problem, therefore, is to minimize the probability of a classification error when the probability distribution function $F(\omega, x)$ is unknown but the data (1.7) are given.

For simplicity consider the two-class classification problem (i.e., $\omega \in \{0, 1\}$) where we use the simplest loss function (1.5).

Thus, the problem of pattern recognition has been reduced to the problem of minimizing the risk on the basis of empirical data. The special feature of this problem is that the set of loss functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is not arbitrary as in the general case described in Section 1.2. The following restrictions are imposed:

- The vector z consists of $n+1$ coordinates: coordinate ω, which takes on only a finite number of values (two values for a two classes problem), and n coordinates $x^1, ..., x^n$ which form the vector x.
- The set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is given by

$$Q(z, \alpha) = L(\omega, \phi(x, \alpha)), \qquad \alpha \in \Lambda$$

  and also takes on only a finite number of values (zero and one for the simplest loss function).

This specific feature of the risk minimization problem characterizes the pattern recognition problem. The problem of pattern recognition forms the simplest learning problem because it deals with the simplest loss function. The loss function in the pattern recognition problem describes a set of *indicator functions*—that is, functions that take only two values, zero and one.

1.4 THE PROBLEM OF REGRESSION ESTIMATION

Two sets of elements X and Y are connected by a functional dependence if to each element $x \in X$ there corresponds a unique element $y \in Y$. This relationship is called a function if X is a set of vectors and Y is a set of scalars.

However, there exist relationships (stochastic dependencies) where to each vector x there corresponds a number y which we obtain as a result of random trials. For each x, let a distribution $F(y|x)$ be defined on Y according to which the selection of the value of y is implemented. The function of the conditional probability expresses the stochastic relationship between y and x.

Now, let the vectors x appear randomly and independently in accordance with a distribution $F(x)$. Then, in accordance with $F(y|x)$, the values of y are realized in random trials. In this case, there exists a joint distribution function $F(x, y)$. In accordance with this measure the observed pairs

$$(y_1, x_1), \ldots, (y_\ell, x_\ell)$$

are formed randomly and independently. Estimating the stochastic dependence based on this empirical data means estimating the conditional distribution function $F(y|x)$, and this is indeed quite a difficult problem. As we show, it leads to the need to solve so-called ill-posed problems.

However, the knowledge of the function $F(y|x)$ is often not required; it is sufficient to determine one of its characteristics, for example the function of conditional mathematical expectation:

$$r(x) = \int y \, dF(y|x). \tag{1.8}$$

This function is called the *regression*, and the problem of its estimation in the set of functions $f(x, \alpha), \alpha \in \Lambda$, is referred to as the problem of regression estimation. We now show that under conditions

$$\int y^2 \, dF(y, x) < \infty, \qquad \int r^2(x) \, dF(y, x) < \infty$$

the problem of regression estimation is reduced to the model of minimizing risk based on empirical data.

Indeed, on the set $f(x, \alpha), \alpha \in \Lambda$ $(f(x, \alpha) \in L_2(P))$, the minimum of the functional

$$R(\alpha) = \int (y - f(x, \alpha))^2 \, dF(y, x) \tag{1.9}$$

(provided the minimum exists) is attained at the regression function if the regression $r(x)$ belongs to $f(x, \alpha), \alpha \in \Lambda$. The minimum of this functional is attained at the function $f(x, \alpha^*)$, which is the closest to regression $r(x)$ in the

metric $L_2(P)$

$$\rho(f_1, f_2) = \sqrt{\int (f_1(x) - f_2(x))^2 \, dF(x)}$$

if the regression $r(x)$ does not belong to the set $f(x, \alpha), \alpha \in \Lambda$.

To show this, denote

$$\Delta f(x, \alpha) = f(x, \alpha) - r(x).$$

Then functional (1.9) can be written in the form

$$R(\alpha) = \int (y - r(x))^2 \, dF(y, x) + \int (\Delta f(x, \alpha))^2 \, dF(y, x) - 2 \int \Delta f(x, \alpha)(y - r(x)) \, dF(y, x).$$

In this expression, the third summand is zero, since according to (1.8)

$$\int \Delta f(x, \alpha)(y - r(x)) \, dF(y, x) = \int \Delta f(x, \alpha) \left[\int (y - r(x)) \, dF(y|x) \right] dF(x) = 0.$$

Thus we have verified that

$$R(\alpha) = \int (y - r(x))^2 \, dF(y, x) + \int (f(x, \alpha) - r(x))^2 \, dF(x).$$

Since the first summand does not depend on α, the function $f(x, \alpha_0)$, which minimizes the risk functional $R(\alpha)$, is the regression if $r(x) \in f(x, \alpha)$, or the function $f(x, \alpha_0)$ which minimizes the risk functional $R(\alpha)$ is the closest function to the regression (in the metric $L_2(P)$), if $r(x)$ does not belong to $f(x, \alpha)$.

This equation also implies that if the regression function $r(x) = f(x, \alpha_0)$ belongs to the given set of functions $f(x, \alpha), \alpha \in \Lambda$, and if for some function $f(x, \alpha^*)$ the risk functional $R(\alpha^*)$ is ε-close to the minimal one

$$R(\alpha^*) - \inf_{\alpha \in \Lambda} R(\alpha) < \varepsilon,$$

then the function $f(x, \alpha^*)$ is $\sqrt{\varepsilon}$-close to the regression in the metric $L_2(P)$:

$$\rho(f(x, \alpha^*), r(x)) = \sqrt{\int (f(x, \alpha^*) - r(x))^2 \, dF(x)} < \sqrt{\varepsilon}.$$

Thus, the problem of estimating the regression may be also reduced to the scheme of minimizing expected risk. The specific feature of this problem is that the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is subject to the following restrictions:

- The vector z consists of $n+1$ coordinates: the coordinate y and n coordinates $x^1, ..., x^n$ forming the vector x. However, in contrast to the pattern recognition problem, the coordinate y as well as the function $f(x, \alpha)$ may take any value in the interval $(-\infty, \infty)$
- The set of loss functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is of the form

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

The important feature of the regression estimation problem is that the loss-function $Q(z, \alpha)$ can take on arbitrary non-negative values whereas in pattern recognition problem it can take only two values.

1.5 PROBLEM OF INTERPRETING RESULTS OF INDIRECT MEASURING

Along with the problem of regression estimation we consider the problem of estimating functional dependencies from indirect measuring.

Suppose one would like to estimate a function $f(t)$ that can be measured at no point of t. At the same time, another function $F(x)$ which is connected with $f(t)$ by operator

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

may admit measurements. It is then required on the basis of measurements (with errors ξ)

$$y_1, ..., y_\ell, \qquad y_i = F(x_i) + \xi_i$$

of function $F(x)$ at points $x_1, ..., x_\ell$ to obtain in a set $f(t, \alpha)$ the solution of the equation. This problem is called the problem of *interpreting results of indirect measurements*.

The formation of the problem is as follows: Given a continuous operator A which maps in 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 in a set of functions $f(t, \alpha)$, $\alpha \in \Lambda$, provided that the function $F(x)$ is unknown, but measurements $y_1, ..., y_\ell$ are given.

We assume that the measuring $F(x)$ does not involve systematic error, that is,

$$Ey_{x_i} = F(x_i)$$

and the random variables y_{x_i} and y_{x_j} $(i \neq j)$ are independent. We also assume that function is defined on the interval $[a, b]$. The points x at which measurements of the function $F(x)$ are carried out are randomly and independently distributed on $[a, b]$ according to uniform distribution.†

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

$$R(\alpha) = \int (y - Af(t, \alpha))^2 p(y|x)\, dy\, dx$$

Using the same decomposition technique as in the previous section we obtain

$$\begin{aligned} R(\alpha) &= \int (y - F(x, \alpha))^2 p(y|x)\, dy\, dx \\ &= \int (y - Af(t))^2 p(y|x)\, dy\, dx + \int (F(x, \alpha) - F(x))^2\, dx \end{aligned}$$

where $f(t)$ and $F(x)$ are the solution of integral equation and its image in E_2 space.

We have thus again arrived at a setup for minimizing expected risk on the basis of empirical data. To solve this problem, we have to find function $f(t, \alpha_0)$, the image of which is the regression function in E_2 space.

- The vector z consists of $n+1$ coordinates: the coordinate y and n coordinates $x^1, \ldots, x^n$ forming the vector x.
- The set of loss-functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is of the form

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

The specific feature of interpreting results of indirect experiments that the problem of solving operator equation

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

may be ill-posed (we will discuss this problem below). In this case not all good approximations to the regression $F(x)$ imply good approximations to the desired solution $f(t)$. In order to approximate the solution of the operator equation well, one has to choose the function that not only provides a small value to the risk functional, but also satisfies some additional constraints that we will discuss later.

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

1.6 THE PROBLEM OF DENSITY ESTIMATION (THE FISHER–WALD SETTING)

Let $p(x, \alpha), \alpha \in \Lambda$, be a set of probability densities containing the required density

$$p(x, \alpha_0) = \frac{dF(x)}{dx}.$$

Consider the functional

$$R(\alpha) = -\int \ln p(x, \alpha)\, dF(x)\,. \tag{1.10}$$

Below we show that:

1. The minimum of the functional (1.10) (if it exists) is attained at the functions $p(x, \alpha^*)$ which may differ from $p(x, \alpha_0)$ only on a set of zero measure.
2. The *Bretagnolle–Huber inequality*

$$\int |\, p(x, \alpha) - p(x, \alpha_0)\,|\, dx \le 2\sqrt{1 - \exp\{R(\alpha_0) - R(\alpha)\}} \tag{1.11}$$

is valid.

Therefore, the functions $p(x, \alpha^*)$ which are ε-close to the minimum

$$R(\alpha^*) - \inf_{\alpha \in \Lambda} R(\alpha) < \varepsilon$$

will be $2\sqrt{1 - \exp\{-\varepsilon\}}$-close to the required density in the metric L_1.

The proof of the first assertion is based on the *Jensen inequality*, which states that for a concave function ψ the inequality

$$\int \psi\,(\Phi(x))\, dF(x) \le \psi\left(\int \Phi(x)\, dF(x)\right) \tag{1.12}$$

is valid.

Consider the functions

$$\psi(u) = \ln u, \qquad \Phi(x) = \frac{p(x, \alpha)}{p(x, \alpha_0)}.$$

Jensen's inequality implies

$$\int \ln \frac{p(x, \alpha)}{p(x, \alpha_0)}\, dF(x) \le \ln \int \frac{p(x, \alpha)}{p(x, \alpha_0)} p(x, \alpha_0)\, dx = \ln 1 = 0.$$

So, the inequality

$$\int \ln \frac{p(x,\alpha)}{p(x,\alpha_0)}\, dF(x) = \int \ln p(x,\alpha)\, dF(x) - \int \ln p(x,\alpha_0)\, dF(x) \le 0$$

is valid. Taking into account the sign in front of the integral (1.10), this inequality proves our first assertion.

To prove the Bretagnolle–Huber inequality, use the following identity:

$$\int p(x,\alpha_0) \ln \frac{p(x,\alpha)}{p(x,\alpha_0)}\, dx = \int p(x,\alpha_0) \ln \left[\min \left(\frac{p(x,\alpha)}{p(x,\alpha_0)},\ 1 \right) \right] dx + \int p(x,\alpha_0) \ln \left[\max \left(\frac{p(x,\alpha)}{p(x,\alpha_0)},\ 1 \right) \right] dx.$$

We apply Jensen's inequality to both terms on the right-hand side of this equality

$$\int p(x,\alpha_0) \ln \frac{p(x,\alpha)}{p(x,\alpha_0)}\, dx \le \ln \int \min\left(p(x,\alpha),\ p(x,\alpha_0)\right) dx + \ln \int \max\left(p(x,\alpha),\ p(x,\alpha_0)\right) dx. \quad (1.13)$$

Note that the following identities are true:

$$\begin{aligned} \min(a,b) &= \frac{a+b-|a-b|}{2}, \\ \max(a,b) &= \frac{a+b+|a-b|}{2}. \end{aligned} \quad (1.14)$$

Substituting (1.14) into (1.13), we obtain

$$\begin{aligned} &\int p(x,\alpha_0) \ln \frac{p(x,\alpha)}{p(x,\alpha_0)}\, dx \\ &\le \ln \left\{ \left(1 - \frac{1}{2} \int |p(x,\alpha) - p(x,\alpha_0)|\, dx \right) \left(1 + \frac{1}{2} \int |p(x,\alpha) - p(x,\alpha_0)|\, dx \right) \right\} \\ &= \ln \left(1 - \left(\frac{1}{2} \int |p(x,\alpha) - p(x,\alpha_0)|\, dx \right)^2 \right). \end{aligned} \quad (1.15)$$

This inequality implies Bretagnolle–Huber inequality.

Thus, the problem of estimating the density in L_1 is reduced to the minimization of the functional (1.10) on the basis of empirical data. We call this setting of the density estimation problem the Fisher–Wald's setting. (In Section 1.8 we consider another setting of this problem.)

The special feature of the density estimation problem in the Fisher–Wald setting is that the set of functions $Q(z,\alpha)$ is subject to the following restrictions:

- The vector z coincides with the vector x.
- The set of functions $Q(z, \alpha), \alpha \in \Lambda$, is of the form

$$Q(z, \alpha) = -\log p(x, \alpha),$$

where $p(x, \alpha)$ is a set of density functions. The loss function $Q(z, \alpha)$ takes on arbitrary values on the interval $(-\infty, \infty)$, whereas in the regression estimation problem it takes on only nonnegative values.

We will restrict our analysis to these three problems. However, many other problems of estimating empirical dependencies can be reduced to the model of risk minimization based on empirical data.

1.7 INDUCTION PRINCIPLES FOR MINIMIZING THE RISK FUNCTIONAL ON THE BASIS OF EMPIRICAL DATA

In the previous sections, we considered the problem of minimizing the risk functional on the basis of empirical data. It was shown that different problems such as pattern recognition, regression estimation, and density estimation can be reduced to this scheme by specifying a loss function in the risk functional.

Now a main question arises:

How can we minimize the risk functional?

We cannot minimize the functional directly since the probability distribution function $F(x)$ that defines the risk is unknown. What shall we do instead? The answer to this question determines an *induction principle* for solving learning problems.

In this book, two induction principles will be considered: (1) the classical one which we introduce in this section and (2) a new one which we consider in Chapter 6.

Principle of Empirical Risk Minimization. Let us, instead of minimizing the risk functional

$$R(\alpha) = \int Q(z, \alpha)\, dF(z), \qquad \alpha \in \Lambda,$$

minimize the functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha), \qquad \alpha \in \Lambda, \tag{1.16}$$

which we call the *empirical risk functional*. The empirical risk functional is constructed on the basis of data

$$z_1, \ldots z_\ell$$

obtained according to distribution function $F(z)$. This functional is defined in explicit form, and it is subject to minimization.

Let the minimum of the risk functional be attained at $Q(z, \alpha_0)$ and let the minimum of the empirical risk functional be attained at $Q(z, \alpha_\ell)$. We shall consider the function $Q(z, \alpha_\ell)$ as an approximation to the function $Q(z, \alpha_0)$. This principle of solving the risk minimization problem is called the *empirical risk minimization (induction) principle.*

The study of this principle is one of the main subjects of this book. The problem is to establish conditions under which the obtained function $Q(z, \alpha_\ell)$ is close to the desired one, $Q(z, \alpha_0)$.

1.8 CLASSICAL METHODS FOR SOLVING FUNCTION ESTIMATION PROBLEMS

Below we show that classical methods for solving our three statistical problems (pattern recognition, regression estimation, and density estimation) are implementations of the principle of empirical risk minimization.

Method of Minimizing Number of Training Error. In Section 1.3 we showed that the minimization using empirical data (training data)

$$(\omega_1, x_1), \ldots, (\omega_\ell, x_\ell)$$

of the risk functional

$$R(\alpha) = \int L(\omega, \phi(x, \alpha))\, dF(\omega, x), \qquad \alpha \in \Lambda$$

on a set of functions $\phi(x, \alpha), \alpha \in \Lambda$, that take on only a finite number of values renders the pattern recognition problem.

Consider the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(\omega_i, \phi(x_i, \alpha)), \qquad \alpha \in \Lambda.$$

In the case when $L(\omega_i, \phi) \in \{0, 1\}$ (0 if $\omega = \phi$ and 1 if $\omega \neq \phi$), minimization of the empirical risk functional produced a function which has the smallest number of errors on the training data.

Least Squares Method for the Regression Estimation Problem. In Section 1.4, we considered the problem of regression estimation as the problem of minimization of the functional

$$R(\alpha) = \int (y - f(x,\alpha))^2 \, dF(y,x), \qquad \alpha \in \Lambda$$

on the set of functions $f(x,\alpha)$, $\alpha \in \Lambda$, on the basis of empirical data

$$(y_1, x_1), \ldots, (y_\ell, x_\ell).$$

For this functional, the empirical risk functional is

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(x_i,\alpha))^2, \qquad \alpha \in \Lambda.$$

According to the empirical risk minimization principle, to estimate the regression function we have to minimize this functional. In statistics, the method of minimizing this functional is known as the "least-squares method."

Maximum Likelihood Method for Density Estimation. In Section 1.5, we considered the problem of density estimation as the problem of minimization of the functional

$$R(\alpha) = -\int \ln p(x,\alpha) \, dF(x), \qquad \alpha \in \Lambda$$

on the set of densities $p(x,\alpha)$, $\alpha \in \Lambda$, using independent identically distributed data

$$x_1, \ldots, x_\ell.$$

For this functional, the empirical risk functional is

$$R_{\text{emp}}(\alpha) = -\sum_{i=1}^{\ell} \ln p(x_i,\alpha).$$

According to the principle of empirical risk minimization, the minimum of this functional provides an approximation of the density. It is the same solution which comes from the maximum likelihood method. (In the maximum likelihood method, a plus sign is used in front of the sum instead of a minus.)

Thus, we find that the classical methods of solving our statistical problems are realizations of the general induction principle of minimizing empirical risk. In subsequent chapters, we will study the general methods of minimizing the risk functionals and then apply them to our specific problems. But before that, we will consider a second approach to the learning problems, which is not based on the scheme of minimizing the risk functional from empirical data.

1.9 IDENTIFICATION OF STOCHASTIC OBJECTS: ESTIMATION OF THE DENSITIES AND CONDITIONAL DENSITIES

1.9.1 Problem of Density Estimation. Direct Setting

Consider methods for identifying stochastic objects. We start with the problem of density estimation. Let ξ be a random variable. The probability of random event

$$F(x) = P\{\xi < x\}$$

is called a *probability distribution function* of the random variable ξ. A random vector $\bar{\xi}$ is a generalization of the notion of a random variable. The function

$$F(\bar{x}) = P\{\bar{\xi} < \bar{x}\},$$

where the inequality is interpreted coordinatewise, is called a *probability distribution function of the random vector* $\bar{\xi}$.

We say that the random variable ξ (random vector $\bar{\xi}$) has a density if there exists a nonnegative function $p(u)$ such that for all x the equality

$$F(x) = \int_{-\infty}^{x} p(u)\,du$$

is valid.

The function $p(x)$ is called a *probability density* of the random variable (random vector). So, by definition, to estimate a probability density from the data we need to obtain a solution of the integral equation

$$\int_{-\infty}^{x} p(u, \alpha)\,du = F(x) \tag{1.17}$$

on a given set of densities $p(x, \alpha)$, $\alpha \in \Lambda$, under conditions that the distribution function $F(x)$ is unknown and a random independent sample

$$x_1, \ldots, x_\ell, \tag{1.18}$$

obtained in accordance with $F(x)$, is given.

One can construct approximations to the distribution function $F(x)$ using the data (1.18)—for example, the so called *empirical distribution function* (1.18) (see Fig. 1.2):

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i), \tag{1.19}$$

where we define for vector[†] u the step function

$$\theta(u) = \begin{cases} 1 & \text{all coordinates of the vector } u \text{ are positive,} \\ 0 & \text{otherwise.} \end{cases}$$

[†] Including scalars as one-dimensional vectors.

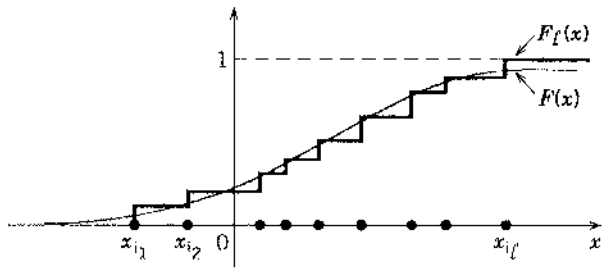

FIGURE 1.2. The empirical distribution function $F_\ell(x)$, constructed from the data $x_1, \ldots, x_\ell$, approximates the probability distribution function $F(x)$.

In the next section, we will show that empirical distribution function $F_\ell(x)$ is a good approximation to the actual distribution function $F(x)$.

Thus, the problem of density estimation is to find an approximation to the solution of the integral equation (1.17) if the probability distribution function is unknown; however, an approximation to this function can be defined.

We call this setting of the density estimation problem *direct setting* because it based on the definition of density. In the following sections we shall discuss the problem of solving integral equations with an approximate right-hand side, but now we turn to a direct setting of the problem of estimating the conditional probability. Using the conditional probability, one can easily solve the pattern recognition problem.

1.9.2 Problem of Conditional Probability Estimation

Consider pairs (ω, x), where x is a vector and ω is a scalar which takes on only k values $\{0, 1, .., k-1\}$. According to the definition, the conditional probability $P(\omega|x)$ is a solution of the integral equation

$$\int_{-\infty}^{x} P(\omega|t)\, dF(t) = F(\omega, x), \tag{1.20}$$

where $F(x)$ is the distribution function of random vectors x, and $F(\omega, x)$ is the joint distribution function of pairs (ω, x).

The problem of estimating conditional probability in the set of functions $P_\alpha(\omega|x)$, $\alpha \in \Lambda$, is to obtain an approximation to the solution of the integral equation (1.20) when both distribution functions $F(x)$ and $F(\omega, x)$ are unknown but the data

$$(\omega_1, x_1), \ldots, (\omega_\ell, x_\ell)$$

is given. As in the case of density estimation, we can approximate the unknown distribution functions $F(x)$ and $F(\omega, x)$ by the empirical distribution

functions (1.19) and function

$$F_\ell(\omega, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i)\delta(\omega, x_i),$$

where

$$\delta(\omega, x) = \begin{cases} 1 & \text{if the vector } x \text{ belongs to the class } \omega, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the problem is to obtain an approximation to the solution of integral equation (1.20) in the set of functions $P_\alpha(\omega|x)$, $\alpha \in \Lambda$, when probability distribution functions $F(x)$ and $F(\omega, x)$ are unknown, but approximations $F_\ell(x)$ and $F_\ell(\omega, x)$ are given.

Note that estimation of the conditional probability function $F(\omega|x)$ is a stronger solution to the pattern recognition problem than the one considered in Section 1.3. In Section 1.3, the goal was to find the best decision rule from the *given set of decision rules*; it did not matter whether this set did or did not contain a good approximation to the supervisor's decision rule. In this statement of the identification problem, the goal is to find the best approximation to the supervisor's decision rule (which is the conditional probability function according to the statement of the problem). Of course, if the supervisor's operator $F(\omega|x)$ is known, then one can easily construct the optimal decision rule. For the case where $\omega \in \{0, 1\}$ and a priori probability of classes are equal, it has the form

$$f(x) = \theta(P(\omega = 1|x) - \tfrac{1}{2}).$$

This is the so-called Bayes rule; it assigns vector x to class 1 if the probability that this vector belongs to the first class is larger than $1/2$ and assigns 0 otherwise. However, the knowledge of the conditional probability not only gives the best solution to the pattern recognition problem, but also provides an estimate of the error probability for any specific vector x.

1.9.3 Problem of Conditional Density Estimation

Finally, consider the problem of conditional density estimation. In the pairs (y, x), let the variables y be scalars and let x be vectors. Consider the equality

$$\int_{-\infty}^{y} \int_{-\infty}^{x} p(t|u)\, dF(u)\, dt = F(y, x), \tag{1.21}$$

where $F(x)$ is a probability distribution function which has a density, $p(y|x)$

is the conditional density of y given x, and $F(y,x)$ is the joint probability distribution function[†] defined on the pairs (y,x).

As before, we are looking for an approximation to the conditional density $p(y|x)$ by solving the integral equation (1.21) on the given set of functions when both distribution functions $F(x)$ and $F(y,x)$ are unknown; and the random, i.i.d. pairs

$$(y_1,x_1),...,(y_\ell,x_\ell) \tag{1.22}$$

are given. As before, we can approximate the empirical distribution function $F_\ell(x)$ and empirical distribution function

$$F_\ell(y,x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(y-y_i)\theta(x-x_i).$$

Thus, our problem is to get an approximation to the solution of the integral equation (1.21) in the set of functions $p_\alpha(y|x), \alpha \in \Lambda$, when the probability distribution functions are unknown but we can construct the approximations $F_\ell(x)$ and $F_\ell(y,x)$ using data (1.22).

Note that the conditional density $p(y|x)$ contains much more information about the behavior of the random value y for fixed x than the regression function. The regression function can be easily obtained from conditional density (see the definition of the regression function (1.8)).

1.10 THE PROBLEM OF SOLVING AN APPROXIMATELY DETERMINED INTEGRAL EQUATION

All three problems of stochastic dependencies estimation can be described in the following general way. It is necessary to solve a linear continuous operator equation

$$Af(\cdot) = F, \qquad f \in \mathcal{F} \tag{1.23}$$

if some functions which form the equation are unknown, but data are given. Using these data the approximations to the unknown functions can be obtained. Let $F_\ell(x)$ and $F_\ell(y,x)$ be approximations to the distribution functions $F(x)$ and $F(y,x)$ obtained from the data.

A difference exists between the problem of density estimation and the problems of conditional probability and conditional density estimation. In the problem of density estimation, instead of an accurate right-hand side of the

[†] Actually, the solution of this equation is the definition of conditional density. Suppose that $p(x)$ and $p(y,x)$ are the densities corresponding to probability distribution functions $F(x)$ and $F(y,x)$. Then equality (1.21) is equivalent to the equality $p(y|x)p(x) = p(y,x)$.

equation we have its approximation. We would like to get an approximation to the solution of Eq. (1.23) from the relationship

$$Af \approx F_\ell, \qquad f \in \mathcal{F}.$$

In the problems of conditional probability and conditional density estimation, not only the right-hand side of Eq. (1.23) is known approximately, but the operator A is known approximately as well (in the left-hand side of integral equations (1.20) and (1.21), instead of the distribution functions, we use their approximations). So our problem is to get an approximation to the solution of Eq. (1.23) from the relationship

$$A_\ell f \approx F_\ell, \qquad f \in \mathcal{F},$$

where A_ℓ is an approximation of the operator A.

The good news about solving these problems is that the empirical distribution function forms a good approximation to the unknown distribution function. In the next section we show that as the number of observations tends to infinity, the empirical distribution function converges to the desired one. Moreover, we shall give an asymptotically exact rate of the convergence for different metrics determining different definitions of a distance between functions.

The bad news is that the problem of solving operator equation (1.23) is the so-called *ill-posed* problem. In Section 1.12 we shall define the concept of "ill-posed" problems and describe the difficulties that arise when one needs to solve ill-posed problems. In the appendix to this chapter we provide the classical theory of solving ill-posed problems which is generalized in Chapter 7 to the case of stochastic ill-posed problems. The theory of solving stochastic ill-posed problems will be used for solving our integral equations.

1.11 GLIVENKO–CANTELLI THEOREM

In the 1930s Glivenko and Cantelli proved one of the most important theorems in statistics. They proved that when the number of observations tends to infinity, the empirical distribution function $F_\ell(x)$ converges to the actual distribution function $F(x)$.

This theorem and its generalizations play an important part both in learning theory and in foundations of theoretical statistics. To discuss this theorem and results related to it accurately, we need to introduce some general concepts which describe the convergence of a stochastic variable.

1.11.1 Convergence in Probability and Almost Sure Convergence

Note that an empirical distribution function is a random function because it is formed on the basis of a random sample of observations. To discuss the problem of convergence of this function we need to measure distance between the empirical distribution function and the actual one. To measure the distance between two functions, different metrics are used. In this book we use three of them: the uniform metric C

$$\rho(g_1(x), g_2(x)) = \sup_x |g_1(x) - g_2(x)|,$$

$L_2(F)$ metric

$$\rho(g_1(x), g_2(x)) = \sqrt{\int (g_1(x) - g_2(x))^2 \, dF(x)},$$

and $L_1(F)$ metric

$$\rho(g_1(x), g_2(x)) = \int |g_1(x) - g_2(x)| \, dF(x).$$

In the case when we measure the distance between random functions $F_\ell(x)$ and some fixed function $F(x)$, random variables

$$a_\ell = a_\ell(x_1, ..., x_\ell) = \rho(F(x), F_\ell(x))$$

are considered. Consider a sequence of random variables

$$a_1, ..., a_\ell, \ldots .$$

We say that a sequence of random variables a_ℓ converges to a random variable a_0 *in probability* if for any $\delta > 0$ the relation

$$P\{|a_\ell - a_0| > \delta\} \underset{\ell \to \infty}{\longrightarrow} 0 \tag{1.24}$$

is valid.

We say also that a sequence of random variables a_n converges to the random variable a_0 *almost surely* (with probability 1) if for any $\delta > 0$ the relation

$$P\{\sup_{\ell > n} | a_\ell - a_0 | > \delta\} \underset{n \to \infty}{\longrightarrow} 0 \tag{1.25}$$

is valid.

It is easy to see that the convergence (1.25) implies the convergence (1.24) which is a weaker mode of convergence. Generally, the convergence (1.24) does not imply the convergence (1.25).

The following classical lemma provides conditions under which convergence in probability implies almost sure convergence (Shiryayev, 1984).

Let $A_1, \ldots, A_n, \ldots$ be a sequence of events.[†] Denote by

$$A = \overline{\lim_{n \to \infty}} A_n$$

the event that an infinite number of events from $A_1, \ldots, A_n, \ldots$ have occurred.

Lemma 1.1 (Borel–Cantelli). (a) *If*

$$\sum_{n=1}^{\infty} P\{A_n\} < \infty,$$

then

$$P\{\overline{\lim_{n \to \infty}} A_n\} = 0.$$

(b) *If*

$$\sum_{n=1}^{\infty} P\{A_n\} = \infty$$

and $A_1, \ldots, A_n, \ldots$ *is sequence of independent events, then*

$$P\{\overline{\lim_{n \to \infty}} A_n\} = 1.$$

Corollary 1. *In order for a sequence of random variables* a_n *to converge to a random variable* a_0 *almost surely, it is sufficient that for any* $\delta > 0$ *the inequality*

$$\sum_{n=1}^{\infty} P\{ \mid a_n - a_0 \mid > \delta\} < \infty$$

be fulfilled.

This inequality forms necessary conditions if a_n is a sequence of independent random variables.

Corollary 2. *Let* ε_n, $n = 1, \ldots,$ *be a sequence of positive values such that* $\varepsilon_n \to 0$ *when* $n \to \infty$. *Then if*

$$\sum_{n=1}^{\infty} P\{ \mid a_n - a_0 \mid > \varepsilon_n\} < \infty,$$

the random variables a_n *converge to a random variable* a_0 *almost surely.*

[†] See Chapter 2 for definition of events.

Convergence in probability will be denoted by

$$a_\ell \xrightarrow[\ell\to\infty]{P} a_0 .$$

Almost sure convergence will be denoted by

$$a_\ell \xrightarrow[\ell\to\infty]{a.s.} a_0.$$

1.11.2 Glivenko–Cantelli Theorem

Now we can formulate the Glivenko–Cantelli theorem.

Theorem 1.1 (Glivenko–Cantelli). *The convergence*

$$\sup_x |F(x) - F_\ell(x)| \xrightarrow[\ell\to\infty]{P} 0$$

takes place.

In this formulation, the Glivenko Cantelli theorem asserts the convergence in probability,[1] in the uniform metric, of the empirical distribution function $F_\ell(x)$ to the actual distribution function $F(x)$.

We will not prove this theorem here, which was proved originally for the one-dimensional case. This theorem and its generalization for the multidimensional case will be derived from the more general assertion, which we shall prove in Chapter 4.

As soon as this theorem has been proved, the problem of the rate of convergence $F_\ell(x)$ to $F(x)$ emerged.

1.11.3 Three Important Statistical Laws

Investigations of the rate of convergence of $F_\ell(x)$ to $F(x)$ for one-dimensional continuous functions $F(x)$ resulted in the establishment of several laws of statistics, in particular the following three:

1. *Kolmogorov–Smirnov Distribution.* The random variable

$$\xi_\ell = \sqrt{\ell} \sup_x | F(x) - F_\ell(x) |$$

has the following limiting probability distribution (Kolmogorov):

$$\lim_{\ell\to\infty} P\{\sqrt{\ell} \sup_x | F(x) - F_\ell(x)| < \varepsilon\} = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1} e^{-2\varepsilon^2 k^2}. \tag{1.26}$$

[1] Below we will see that almost sure convergence takes place as well.

The random variables

$$\xi_\ell^+ = \sqrt{\ell}\sup_x(F(x) - F_\ell(x))\,,$$

$$\xi_\ell^- = \sqrt{\ell}\sup_x(F_\ell(x) - F(x))$$

have the following limiting probability distributions (Smirnov):

$$\begin{aligned}\lim_{\ell\to\infty} P\{\sqrt{\ell}\sup_x(F(x) - F_\ell(x)) < \varepsilon\} &= 1 - e^{-2\varepsilon^2},\\ \lim_{\ell\to\infty} P\{\sqrt{\ell}\sup_x(F_\ell(x) - F(x)) < \varepsilon\} &= 1 - e^{-2\varepsilon^2}.\end{aligned} \tag{1.27}$$

2. *The Law of the Iterated Logarithm.* The equality

$$P\left\{\lim_{n\to\infty}\sup_{\ell>n}\sup_x\sqrt{\frac{2\ell}{\ln\ln\ell}}\,|\,F(x) - F_\ell(x)\,| = 1\right\} = 1 \tag{1.28}$$

holds true.

3. *Smirnov Distribution.* The statistic

$$\omega^2 = \ell\int(F(x) - F_\ell(x))^2\,dF(x)$$

(the so-called *omega square statistic*) has the limiting distribution

$$\begin{aligned}&\lim_{\ell\to\infty} P\{\ell\int(F(x) - F_\ell(x))^2\,dF(x) < \varepsilon\}\\ &\quad = 1 - \frac{2}{\pi}\sum_{k=1}^{\infty}\int_{(2k-1)\pi}^{2k\pi}\frac{\exp\{-\dfrac{\lambda^2\varepsilon}{2}\}}{\sqrt{-\lambda\sin\lambda}}\,d\lambda\,.\end{aligned}$$

We shall not prove these statistical laws. For our purpose of constructing the learning theory, we need more general laws which we shall derive in Chapters 4 and 5. Now our goal is to use the laws above to estimate the bounds for distribution function $F(x)$ provided the estimate $F_\ell(x)$.

We derive these bounds from the Kolmogorov–Smirnov law (1.27). For this purpose we consider for some η $(0 < \eta < 1)$ the equality

$$1 - e^{-2\varepsilon^2\ell} = 1 - \eta$$

which we solve with respect to ε

$$\varepsilon = \sqrt{-\frac{\ln\eta}{2\ell}}.$$

Now (1.27) can be described as follows: With probability $1 - \eta$ simultaneously for all x the inequalities

$$F_\ell(x) - \sqrt{-\frac{\ln \eta}{2\ell}} \le F(x) \le F_\ell(x) + \sqrt{\frac{\ln \eta}{2\ell}} \tag{1.29}$$

are valid as $\ell \to \infty$.

Similarly, the iterated logarithm law (1.28) implies that when

$$\ell \to \infty$$

simultaneously for all x, the inequalities

$$F_\ell(x) - \sqrt{\frac{\ln \ln \ell}{2\ell}} \le F(x) \le F_\ell(x) + \sqrt{\frac{\ln \ln \ell}{2\ell}}$$

are valid. These inequalities are tight.

To estimate the density we have to solve an integral equation where the right-hand side of the equation is unknown, but approximations which converge to the actual function are given. But even if the approximation $F_\ell(x)$ tends to $F(x)$ with a high asymptotic rate of convergence, the problem of solving our integral equations is hard, since (as we will see in the next section) it is an ill-posed problem.

1.12 ILL-POSED PROBLEMS

We say that the solution of the operator equation

$$Af(t) = F(x) \tag{1.30}$$

is *stable* if a small variation in the right-hand side $F(x) \in F(x, \alpha)$ results in a small change in the solution; that is, if for any ε there exists $\delta(\varepsilon)$ such that the inequality

$$\rho_{E_1}(f(t, \alpha_1),\ f(t, \alpha_2)) \le \varepsilon$$

is valid as long as inequality

$$\rho_{E_2}(F(x, \alpha_1),\ F(x, \alpha_2)) \le \delta(\varepsilon)$$

holds. Here the indices E_1 and E_2 denote that the distance is defined in the metric spaces E_1 and E_2, respectively (the operator equation (1.30) maps functions of space E_1 into functions of space E_2).

We say that the problem of solving the operator equation (1.30) is *well-posed in the Hadamard sense* if the solution of the equation

- *exists,*
- *is unique,* and
- *is stable.*

The problem of solving an operator equation is considered *ill-posed* if the solution of this equation violates at least one of the above-mentioned requirements. In this book, we consider ill-posed problems when the solution of the operator equation exists, is unique, but is not stable.

This book considers ill-posed problems 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 also be valid for equations defined by any other linear continuous operator.

Thus, consider Fredholm's integral equation of type I:

$$\int_0^1 K(t,x)f(t)\,dt = F(x) \tag{1.31}$$

defined by the kernel $K(t,x)$, which is continuous almost everywhere on $0 \leqslant t \leqslant 1,\ 0 \leqslant x \leqslant 1$. This kernel maps the set of functions $\{f(t)\}$, continuous on $[0,1]$, onto the set of functions $\{F(x)\}$ also continuous on $[0,1]$.

We shall now show that the problem of solving the equation (1.31) is an ill-posed one. For this purpose we note that the continuous function $G_\nu(x)$ which is formed by means of the kernel $K(t,x)$:

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

possesses the property

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

Consider the integral equation

$$\int_0^1 K(t,x)f^*(t)\,dt = F(x) + G_\nu(x).$$

Since the Fredholm equation is linear, the solution of this equation has the form

$$f^*(t) = f(t) + \sin \nu t,$$

where $f(t)$ is the solution of Eq. (1.31). For sufficiently large ν, the right-hand side of this equation differs from the right-hand side of (1.31) only by the small amount $G_\nu(x)$, while its solution differs by the amount $\sin \nu t$.

The Fredholm integral equation is the equation we shall consider in this book. Here are some examples of problems connected with a solution of this equation:

Example 1 (The Problem of Identifying Linear Dynamic Systems). It is known that dynamic properties of linear homogeneous objects

$$y(t) = Ax(t)$$

with one output are completely described by the impulse response function $f(\tau)$. The function $f(\tau)$ is the response of the system to a unit impulse $\theta(t)$ served at the system at time $\tau = 0$.

Knowing this function, one can compute the response of the system to the 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 a system is reduced to the determination of the weight function $f(\tau)$.

It is also known that for a linear homogeneous system, the Wiener–Hopf equation

$$\int_0^\infty R_{xx}(t - \tau) f(\tau)\, d\tau = R_{yx}(t) \tag{1.32}$$

is valid.

Equation (1.32) connects the autocorrelation function $R_{xx}(u)$ of a stationary random process at the input of the object with the weight function $f(\tau)$ and the joint correlation function of the input and output signals $R_{yx}(t)$.

Thus, the problem of identifying a linear system involves determining the 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; that is, it is a problem of solving integral equation (1.32) on the basis of empirical data.

Example 2 (The Problem of Estimating Derivatives). Let measurements of a smooth function $F(x)$ at ℓ points of the interval $[0, 1]$ be given. Suppose that the points at which the measurements were taken are distributed randomly and independently according to the uniform distribution. The problem is to estimate the derivative $f(x)$ of the function $F(x)$ on $[0, 1]$.

It is easy to see that the problem is reduced to solving the Volterra integral equation of type 1,

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

under the condition that the ℓ measurements

$$y_1, \ldots, y_\ell$$

of the function $F(x)$ at the points

$$x_1, \ldots, x_\ell$$

are known. Equivalently, it is reduced to the solution of the Fredholm equation of the type I,

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

where

$$\theta(u) = \begin{cases} 1 & \text{if } u > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Note that in the case when $F(x)$ is a monotonically increasing function satisfying the conditions $F(0) = 0$, $F(1) = 1$, we have the problem of density estimation.

In the general case when the kth derivative has to be estimated, the following integral equation has to 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 $F(x)$ the empirical data $y_1, \ldots, y_\ell$ are used. Here $F^{(j)}(0)$ is the value of the jth derivative at zero.

The main difficulty in solving integral equations stems from the fact that this is an ill-posed problem since the solution of the equation is *unstable*. In the mid-1960s, several methods for solving unstable problems of mathematical physics were proposed. In the appendix to this chapter, we shall present the so-called "regularization method" proposed by A. N. Tikhonov. This method is applicable for solving integral equations when instead of knowing the function on the right-hand side of an equation, one knows the sequence of approximations which converges to an unknown function with probability one.

In the 1970s we generalized the theory of the regularization method for solving the so-called stochastic ill-posed problems. We define stochastic ill-posed problems as problems of solving operator equations in the case when approximations of the function on the right-hand side converge in probability to an unknown function and/or when the approximations to the operator converge in probability to an unknown operator. This generalization will be presented in Chapter 7. We show that the regularization method solves stochastic ill-posed problems as well. In particular, it solves our learning problems: estimating densities, conditional densities, and conditional probabilities.

1.13 THE STRUCTURE OF THE LEARNING THEORY

Thus, in this chapter we have considered two approaches to learning problems. The first approach (imitating the supervisor's operator) brought us to the problem of minimizing a risk functional on the basis of empirical data.

The second approach (identifying the supervisor's operator) brought us to the problem of solving some integral equation when the elements of an equation are known only approximately.

It has been shown that the second approach gives more details on the solution of pattern recognition and regression estimation problems.

Why in this case do we need both approaches? As we mentioned in the last section, the second approach, which is based on the solution of the integral equation, forms an ill-posed problem. For ill-posed problems, the best that can be done is to obtain a sequence of approximations to the solution which converges in probability to the desired function when the number of observations tends to infinity. For this approach, there exists no way to evaluate how well the problem can be solved if a finite number of observations is used. In the framework of this approach to the learning problem, any exact assertion is asymptotic.

That is why the first approach, based on minimizing the risk functional from empirical data of the finite size ℓ, may be more apropriate for our purposes.

In the following chapters, we show that in the framework of the first approach one can estimate how close the risk functional of the chosen function is to the smallest possible one (for a given set of functions).

This means that if the function $Q(z, \alpha_\ell)$ has been chosen via an appropriate induction principle (for example, the principle of empirical risk minimization), one can assert that with probability $1 - \eta$ the value of the risk $R(\alpha_\ell)$ for this function does not exceed the smallest possible value of risk $\inf_{\alpha \in \Lambda} R(\alpha)$ (for a given set of functions) by more than ε. Here ε depends only on η, ℓ and one more parameter describing some general properties (capacity) of a given set of functions.

In other words, it will be shown that for algorithms selecting functional dependencies based on empirical risk minimization induction principles, one can guarantee that with probability at least $1 - \eta$ the inequality

$$R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) \leq \varepsilon(\ell, \eta, \cdot) \tag{1.33}$$

holds true.

Recall that for the pattern recognition problem the goal is to obtain the solution for which the value of risk is ε-close to minimal (see Section 1.3).

For the regression estimation problem, the ε-closeness of the risk functional to the minimal one guarantees that the chosen function is $\sqrt{\varepsilon}$-close to the regression function in the $L_2(F)$ metric (see Section 1.4).

For the density estimation problem, the ε-closeness of the risk functionals

to the minimal one implies the $(2\sqrt{1 - \exp\{-\varepsilon\}})$-closeness of approximation to the actual density in the $L_1(F)$ metric (see Section 1.5).

Therefore the main problem in this approach (both theoretical and practical) is to find the method which provides the smallest ε on the right-hand side of inequality (1.33) (for a given number of observations).

To do this well, four levels of the theory should be developed. These are:

1. *Theory of Consistency of the Learning Processes.* The goal of this part of the learning theory is to give a complete description of the conceptual (asymptotic) models of the learning processes—that is, to find the necessary and sufficient conditions of consistency of the learning processes. (Informally, the conditions for convergence to zero of the ε in (1.33) as the number of observations ℓ tends to infinity. The exact definition of consistency is given in Chapter 3.)

   Why do we need this asymptotic (conceptual) part of the theory if our goal is to obtain the best solution for a finite number of observations? The conceptual (asymptotic) part of the learning theory is important since to find the condition for consistency one has to introduce some concepts in terms of which the theory can be developed. For example, the concept which characterizes the capacity of a given set of functions (the dot in arguments of ε in the inequality (1.33)). Generally, it is possible to use several different constructions. However, it is important to develop the theory on the basis of such constructions which are not only sufficient for the consistency of learning process, but are *necessary* as well. This gives us a guarantee that the theory which we develop using these constructions is general and from the conceptual point of view cannot be improved.

2. *Theory of Estimating the Rate of Convergence of the Learning Processes.* This part of the learning theory is devoted to obtaining nonasymptotic bounds on the generalization ability of the learning machines (ε on the right-hand side of inequality (1.33)). We obtain these bounds using the concepts developed in the conceptual part of the theory. In this book, we consider a theory of distribution-free bounds of the rate of convergence (the theory that does not use a priori information about the unknown probability measure). The main requirement of this part of the theory is to find a way to construct bounds for different sets of functions.

3. *Theory for Controlling the Rate of Convergence of the Learning Processes.* The bounds on generalization ability will be used for developing the new induction principles that guarantee the best solution of the learning problem for a given finite set of observations.

   These induction principles are based on the trade-off between complexity of the chosen function (capacity of the set of functions from which the function is chosen) and the value of empirical risk which can be achieved using this function. This trade-off led to some functional different from the empirical risk functional that should be minimized.

Table 1.1. Structure of Learning Theory and Its Representation in this Book

Parts of the Theory	Chapters	Content of the Chapters
1. Theory of consistency of the learning processes	Chapter 3 Chapter 14 Chapter 15 Chapter 16	Review of the theory Proofs of the theorems Proofs of the theorems Proofs of the theorems
2. Theory of bounds	Chapter 4 Chapter 5	For indicator functions For real-valued functions
3. Theory of controlling the generalization	Chapter 6 Chapter 7 Chapter 8	SRM induction principle Stochastic ill-posed problems New setting of the problem
4. Theory of the learning algorithms and its applications	Chapter 9 Chapter 10 Chapter 11 Chapter 12 Chapter 13	Classical approaches SVM for pattern recognition SVM for function estimation Examples of pattern recognition Examples of function estimation

Obtaining these functionals in explicit form is the goal of this part of the theory.

4. *Theory of the Algorithms.* Finally, there is a theory of learning algorithms. The goal of this part of the theory is to develop tools for minimizing the functionals describing the trade-off. In order to minimize these functionals, it is necessary to develop algorithms which can control both the minimization of empirical risk in a given set of functions and the choice of a set of functions with appropriate capacity.

In this book, we consider all parts of the theory of minimization of the risk functional from empirical data.

We consider the theory of solving stochastic ill-posed problem as well, and we apply it to estimate density, conditional density, and conditional probability. This theory describes sufficient conditions for consistency of the solution and, for some cases, the asymptotic rate of convergence of the solution. Of course, the results of asymptotic theory is not enough to guarantee the success if the algorithms use limited samples. In the framework of this theory, our hope is that asymptotic properties established in the theory are also valid for not very large ℓ.

Table 1.1 shows the structure of the learning theory and its representation in this book.

Chapter 2 is not indicated in this table. The content of that chapter goes beyond the learning theory. It, however, is very important for the general understanding of the nature of learning problems. We show in this chapter how deeply these problems are connected with fundamental problems of theoretical statistics.

APPENDIX TO CHAPTER 1: METHODS FOR SOLVING ILL-POSED PROBLEMS

A1.1 THE PROBLEM OF SOLVING AN OPERATOR EQUATION

We say that two sets of elements $f \in \mathcal{M}$ and $F \in \mathcal{N}$ are connected by *functional dependency* if to each element $f \in \mathcal{M}$ there corresponds a unique element $F \subset \mathcal{N}$.

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 $\mathcal{M}$ onto 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

$$Af(t) = F(x) \tag{A1.1}$$

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 $\mathcal{M}$ onto $\mathcal{N}$ and a function $F(x) \in \mathcal{N}$, there exists a unique solution of the operator equation (A.1.1). However, finding 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 $f \subset \mathcal{M}$ belong to a metric space E_1 with metric $\rho_1(\cdot, \cdot)$, and the elements $F \subset \mathcal{N}$ belong to a metric space E_2 with metric $\rho_2(\cdot, \cdot)$. 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 one-to-one operator $\mathcal{M}$ onto $\mathcal{N}$. The solution of such an operator equation exists and is unique, that is, there exists inverse operator A^{-1} from $\mathcal{N}$ onto $\mathcal{M}$:

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

The basic problem is whether the inverse operator is continuous.

If the operator A^{-1} is continuous, then close preimages will correspond to close function in $\mathcal{N}$, that is, the solution of the operator equation (A1.1) will be *stable*.

If, however, the inverse operator is not continuous, then the solution of the operator equation can be *nonstable*. In this case according to Hadamard's definition (Chapter 1, Section 1.12), the problem of solving an operator equation is ill-posed.

It turns out that in many important cases, for example, for a so-called completely continuous operator 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 of the functions in the space E_1 into a compact set of the space E_2—that is, if each bounded infinite sequence in E_1

$$f_1, f_2, \ldots, f_i, \ldots, \qquad \|f_i\| \le c, \tag{A1.2}$$

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

$$Af_1, \ldots, Af_i, \ldots, \tag{A1.3}$$

such that a convergent subsequence

$$Af_{i_1}, \ldots, Af_{i_k}, \ldots \tag{A1.4}$$

can be extracted from it.

We will show that if the space E_1 contains bounded noncompact sets, then the inverse operator A^{-1} for an absolutely continuous operator A need not be continuous.

Indeed, consider a bounded noncompact set in E_1. Select in this set an infinite sequence (A1.2) such that no subsequence of it is convergent. An infinite sequence (A1.3) from which convergent subsequence (A1.4) 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

$$f_{i_1}, \ldots, f_{i_k}, \ldots, \tag{A1.5}$$

would correspond to the sequence (A1.4) in E_1 which is a subsequence of (A1.2). This, however, contradicts the choice of (A1.2).

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

$$Af \equiv \int_a^b K(t,x)f(t)\,dt$$

with the kernel $K(t,x)$ continuous in the domain $a \leq t \leq b$, $a \leq x \leq b$. These operators are completely continuous from $C[a,b]$ into $C[a,b]$. The proof of this fact can be found in textbooks on functional analysis (see, for example, Kolmogorov and Fomin (1970)).

A1.2 PROBLEMS WELL-POSED IN TIKHONOV'S SENSE

Definition. The problem of solving the operator equation

$$Af = F \tag{A1.6}$$

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

1. The solution of (A1.6) exists for each $F \in A\mathcal{M}^* \equiv \mathcal{N}^*$ and belongs to $\mathcal{M}^*$.
2. The solution belonging to $\mathcal{M}^*$ is unique for any $F \in \mathcal{N}^*$.
3. The solutions belonging to $\mathcal{M}^*$ are stable with respect to $F \in \mathcal{N}^*$.

If $\mathcal{M}^* \equiv \mathcal{M}$ and $\mathcal{N}^* \equiv \mathcal{N}$, then correctness in Tikhonov's sense corresponds to correctness in Hadamard's sense. The meaning of Tikhonov'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 to a compact set $\mathcal{M}^*$, then it constitutes a correctness class.

Lemma. *If A is a continuous one-to-one operator defined on a compact set $\mathcal{M}^* \subset \mathcal{M}$, 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 to 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 has been 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 Tikhonov'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 problems. We shall consider one of them.

A1.3 THE REGULARIZATION METHOD

A1.3.1 Idea of Regularization Method

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

Suppose that it is required to solve the operator equation

$$Af = F \tag{A1.7}$$

defined by a continuous one-to-one operator A acting from $\mathcal{M}$ into $\mathcal{N}$. Suppose the solution of (A1.7) exists.

Consider a lower semicontinuous functional $W(f)$, which we shall call the *regularizer* and which possess the following three properties:

1. The solution of the operator equation belongs to the domain of definition $D(W)$ of the functional $W(f)$.
2. On the domain of the definition, functional $W(f)$ admits real-valued nonnegative values.
3. The sets

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

are all compact.

The idea of regularization is to find a solution for (A1.7) 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 Eq. (A1.7) and therefore would also be ill-posed) but is an "improved" functional

$$R_\gamma(\hat{f}, F) = \rho_2^2(A\hat{f}, F) + \gamma W(\hat{f}), \qquad \hat{f} \in D(W) \tag{A1.8}$$

with *regularization parameter* $\gamma > 0$. We will prove that the problem of minimizing the functional (A1.8) is stable, that is, 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 $R_\gamma(f, F_\delta)$.

A1.3.2 Main Theorems About the Regularization Method

The problem in the theory of regularization is to determine a relationship between δ and γ such that the sequence of solutions f_δ^γ of regularized problems $R_\gamma(f, F_\delta)$ converges as $\delta \longrightarrow 0$ to the solution of the operator equation (A1.7).

The following theorem establishes these relations.

Theorem 1. *Let E_1 and E_2 be metric spaces, and suppose for $F \subset \mathcal{N}$ there exists a solution $f \in D(W)$ of Eq. (A1.7). Let instead of an exact right-hand*

side F of Eq. (A1.7), approximations[†] *$F_\delta \in E_2$ be given such that $\rho_2(F, F_\delta) \le \delta$. Suppose the values of parameter γ are chosen in such a manner that*

$$\gamma(\delta) \to 0 \qquad \text{for} \quad \delta \to 0,$$
$$\lim_{\delta \to 0} \frac{\delta^2}{\gamma(\delta)} \le r < \infty. \tag{A1.9}$$

Then the elements $f_\delta^{\gamma(\delta)}$ minimizing the functionals $R_{\gamma(\delta)}(f, F_\delta)$ on $D(W)$ 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 $f^\gamma \in D(W)$ exists which minimizes the functional $R_\gamma(f, F)$ on $D(W)$.

Let γ and δ satisfy the relation (A1.9). Consider a sequence of elements $f_\delta^{\gamma(\delta)}$ minimizing $R_{\gamma(\delta)}(f, F_\delta)$, and show that the convergence

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

is valid.

Indeed, by definition of $f_\delta^{\gamma(\delta)}$ we have

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

Taking into account that

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

we conclude

$$W(f_\delta^{\gamma(\delta)}) \le W(f) + \frac{\delta^2}{\gamma(\delta)},$$

$$\rho_2^2(Af_\delta^{\gamma(\delta)}, F_\delta) \le \gamma(\delta) \left(W(f) + \frac{\delta^2}{\gamma(\delta)} \right)$$

Since the conditions (A1.9) 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 $c^* = W(f) + r + \varepsilon > 0$, $\varepsilon > 0$, and their images $F_\delta^{\gamma(\delta)}$ $Af_\delta^{\gamma(\delta)}$ are convergent:

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

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

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

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

are also converged.

The theorem is thus proved.

In a Hilbert space the functional $W(f)$ 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 $D(W)$ 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 1; namely, γ should converge to zero slower than δ^2.

Thus the following theorem is valid.

Theorem 2. *Let E_1 be a Hilbert space and $W(f) = \|f\|^2$. Then for $\gamma(\delta)$ satisfying the relations (A1.9) 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 $\|f\| \le c$ 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

$$\|f\| \le \liminf_{i\to\infty} \|f_i\|. \tag{A1.10}$$

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

It is not difficult to check 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 (A1.10) the inequality

$$\|f\| \le \liminf_{\delta\to 0} \|f_\delta^{\gamma(\delta)}\|$$

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

$$\limsup_{\delta\to 0} \|f_\delta^{\gamma(\delta)}\|^2 \le \lim_{\delta\to 0} \left(\|f\|^2 + \frac{\delta^2}{\gamma(\delta)} \right) = \|f\|^2.$$

Hence the convergence of the norms is valid:

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

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

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

The theorem is thus proved.

The theorems presented above are fundamentals in regularization theory. Using these theorems the feasibility of solving ill-posed problems is established.

In Chapter 7 we consider the so-called stochastic ill-posed problems and generalize these theorems for the stochastic case. Using the method of regularization for stochastic ill-posed problems we consider our learning problems of estimating densities, conditional probabilities, and conditional densities.