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

V. Vapnik, *Estimation of Dependences Based on Empirical Data*

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. □