Chapter 6

A Method of Minimizing Empirical Risk for the Problem of Pattern Recognition

§1 A Method of Minimizing Empirical Risk

In the preceding three chapters the estimation of dependences was associated with the methods of estimating probability densities. The determination of the function which minimizes the expected risk

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \tag{6.1}$$

on the basis of the empirical data

$$x_1, y_1; \ldots; x_l, y_l \tag{6.2}$$

was reduced to estimating the density $\hat{P}(x, y)$ on the basis of the sample (6.2) and minimization of the functional

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

As was mentioned in Chapter 2, this method of minimizing the risk (6.1) generally is not reasonable, because the problem of density estimation is a more difficult problem than the minimization of the expected risk. Only when a substantial prior information is available about the desired density $P(x, y)$, so that the function $P(x, y)$ can be defined up to its parameters, is this approach plausible. Methods of parametric statistics developed for this particular case were utilized in the preceding chapters.

However, in specific problems the structure of the density $P(x, y)$ is unknown. Thus the successful application of methods of parametric statistics hinges on the assumption that the hypothesized density structure corresponds to the true one.

V. Vapnik, *Estimation of Dependences Based on Empirical Data*
© Springer Science+Business Media, Inc. 2006

Starting with this chapter, we shall study methods of estimating dependences which do not require density estimation. The basis for these methods is the principle of minimizing the empirical risk, according to which as the minimum point of the functional (6.1) one takes the minimum point of the empirical functional

$$I_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2, \tag{6.3}$$

constructed from a random independent sample (6.2). Let the minimum of functional (6.3) be attained for $F(x, \alpha_{\text{emp}})$. The problem is to establish when the obtained function $F(x, \alpha_{\text{emp}})$ is close to the function $F(x, \alpha_0)$ which minimizes (6.1) in $F(x, \alpha)$.

Above (Chapter 2, Section 6) we have associated this problem with the problem of the uniform convergence of the means to their mathematical expectations, i.e., with the situation when for any given value of deviation $\varkappa$ the inequality

$$P\left\{\sup_{\alpha} |I(\alpha) - I_{\text{emp}}(a)| > \varkappa\right\} < \eta \tag{6.4}$$

can be asserted.

Let (6.4) be satisfied. Then the inequality

$$P\{I(\alpha_{\text{emp}}) - I(\alpha_0) > 2\varkappa\} < \eta \tag{6.5}$$

is valid. In other words, if (6.4) holds, then with probability $1 - \eta$ the deviation of the function (solution) $F(x, \alpha_0)$ which is the best in the class $F(x, \alpha)$ from the function which yields a minimum for the empirical risk $F(x, \alpha_{\text{emp}})$ does not exceed $2\varkappa$.

Indeed, the condition (6.4) implies that with probability $1 - \eta$ the two inequalities

$$\begin{aligned} I(\alpha_{\text{emp}}) - I_{\text{emp}}(\alpha_{\text{emp}}) &< \varkappa, \\ I_{\text{emp}}(\alpha_0) - I(\alpha_0) &< \varkappa \end{aligned} \tag{6.6}$$

are simultaneously satisfied. Moreover, since α_{emp} and α_0 are the minimum points of $I_{\text{emp}}(\alpha)$ and $I(\alpha)$, the inequality

$$I_{\text{emp}}(\alpha_{\text{emp}}) \leq I_{\text{emp}}(\alpha_0) \tag{6.7}$$

is valid. The inequalities (6.6) and (6.7) yield that

$$I(\alpha_{\text{emp}}) - I(\alpha_0) < 2\varkappa. \tag{6.8}$$

And since the inequalities (6.6) are both fulfilled simultaneously with probability $1 - \eta$, so is (6.8). Consequently

$$P\{I(\alpha_{\text{emp}}) - I(\alpha_0) > 2\varkappa\} < \eta. \tag{6.9}$$

In this chapter we shall consider the theory of uniform convergence of the means to the mathematical expectations as applied to the problem of pattern recognition: i.e., in the case when the loss function in the functional of expected risk takes only two values, zero and one. In Chapter 7, for the problem of regression estimation we shall extend the results obtained to the case when the loss function takes on an arbitrary form in the interval $(0, \infty)$. It is important to note here that the validity of basic theorems proved in these chapters does not depend on the form of the loss function. Therefore in spite of a quadratic loss function used in the text we shall obtain a general theory of risk minimization.

§2 Uniform Convergence of Frequencies of Events to Their Probabilities

Consider the functional whose minimization is the essence of the pattern recognition problem:

$$I(\alpha) = P(\alpha) = \int (\omega - F(x, \alpha))^2 P(x, \omega)\, dx\, d\omega. \tag{6.10}$$

As has already been mentioned, this functional defines for each decision rule the probability of erroneous classification. The empirical functional

$$I_{\text{emp}}(\alpha) = \nu(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (\omega_i - F(x_i, \alpha))^2, \tag{6.11}$$

computed by means of the sample

$$x_i, \omega_1; \ldots ; x_l, \omega_l, \tag{6.12}$$

defines for each decision rule the frequency of incorrect classification.

According to the classical theorems of probability theory the frequency of occurrence of an event converges to the probability of this event as the number of trials increases indefinitely. Formally this means that for any fixed α and $\varkappa$ the relation

$$\lim_{l \to \infty} P\{|P(\alpha) - \nu(\alpha)| > \varkappa\} = 0 \tag{6.13}$$

holds. However (cf. Chapter 2, Section 6), the condition (6.13) does not imply that the rule which minimizes (6.11) will yield a value of the functional (6.10) close to the minimal. For l sufficiently large the proximity between the solution obtained and the best one does follow from a stronger condition which stipulates that the equality

$$\lim_{l \to \infty} P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\} = 0 \tag{6.14}$$

is valid for any $\varkappa$. In this case we say that the *uniform convergence offrequencies of events to their probabilities over a class of events* $S(\alpha)$ is valid. Each event $S(\alpha^*)$ in the class $S(\alpha)$ is given by the decision rule $F(x, \alpha^*)$ as a set of pairs x, ω for which the equality $(\omega - F(x, \alpha^*))^2 = 1$ is satisfied.

Below we shall present conditions which assure uniform convergence of frequencies of events to their probabilities and at the same time determine the domain of applicability of the method of minimizing empirical risk. However, we first note that application of the method of minimizing the empirical risk does not guarantee a successful solution of the problem of estimating dependences. Here is an example of an algorithm for pattern recognition which minimizes the empirical risk but at the same time one cannot guarantee that the constructed decision rule will be close to the best in a given class: Elements of the sample are stored in memory, and each situation to be recognized is compared with the examples available in memory. If the situation at hand coincides with one of the examples it will be attributed to the class to which the example belongs. If, however no analogous example is available in memory, the situation is attributed to the first class. It is obvious that such a device cannot improve itself, since usually only a negligible fraction of the possible situations will correspond to the sample. At the same time, such a device classifies the elements of the sample without error, i.e., the algorithm minimizes the empirical risk down to zero.

Below we shall verify that this algorithm uses a set of decision rules which form a system of events over which uniform convergence does not hold.

§3 A Particular Case

When does the uniform convergence of frequencies to probabilities take place? Consider the simple case where the class of decision rules $F(x, \alpha)$ is finite, consisting of N rules:

$$F(x, \alpha_1), \dots, F(x, \alpha_N).$$

An event A_i corresponds to each decision rule $F(x, \alpha_i)$ consisting of pairs x, ω such that $(\omega - F(x, \alpha_i))^2 = 1$. This defines a finite number N of events $A_1, \dots, A_N$.

For each fixed event the law of large numbers is valid (the frequency converges to the probability as the number of trials increases indefinitely). One of the specific forms of this law is the Hoeffding inequality:

$$P\{|P(\alpha_i) - \nu(\alpha_i)| > \varkappa\} < 2\exp\{-2\varkappa^2 l\}. \tag{6.15}$$

We are however interested in uniform convergence, i.e., in the probability of simultaneous fulfillment of inequalities

$$|P(\alpha_i) - \nu(\alpha_i)| \le \varkappa, \qquad i = 1, 2, \dots, N.$$

This probability can be easily bounded from above if the probability of occurrence of each one of the inequalities (6.15) is assessed separately:

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} \le \sum_{i=1}^{N} P\{|P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}.$$

Taking into account the inequality (6.15), we obtain

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} < 2N \exp\{-2\varkappa^2 l\}. \tag{6.16}$$

This inequality implies that for a finite number of events the uniform convergence of frequencies of occurrences of events to the corresponding probabilities is always valid, i.e., the limit

$$\lim_{l\to\infty} P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} = 0.$$

We now require that the probability of the realization of the event

$$\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\}$$

not exceed η, i.e., that the inequality

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} < \eta \tag{6.17}$$

will be fulfilled. It follows from the bound (6.16) that the inequality (6.17) is definitely satisfied if the quantities N, l, $\varkappa$, and η are connected by

$$2N \exp\{-2\varkappa^2 l\} = \eta. \tag{6.18}$$

If one solves Equation (6.18) for $\varkappa$, then for given N, l, and η an estimator of the maximal deviation of the frequencies from the corresponding probability in the class of events under consideration is obtained:

$$\varkappa = \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}}. \tag{6.19}$$

If, however we solve Equation (6.18) for l, then we obtain the size of the sample such that with probability at least $1 - \eta$ one can assert that the maximal deviation of the frequency from the probability over this class does net exceed $\varkappa$:

$$l = \frac{\ln N - \ln(\eta/2)}{2\varkappa^2}. \tag{6.20}$$

We have thus proved the following theorem:

Theorem 6.1. *Let the set of decision rules consist of N elements, and for decision rules $F(x, \alpha_i)$ let the frequency of errors in the sample of size l be equal to*

$\nu(\alpha_i)$. Then with probability $1 - \eta$ one may assert that the inequality

$$\nu(\alpha_i) - \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}} \leq P(\alpha_i) \leq \nu(\alpha_i) + \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}}$$

is valid simultaneously for all decision rules.

Remark. Since the inequalities are valid for all N rules, Theorem 6.1 determines a confidence interval for the quality of a decision rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk among N rules. This interval is

$$\nu(\alpha_{\text{emp}}) - \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}} \leq P(\alpha_{\text{emp}}) \leq \nu(\alpha_{\text{emp}}) + \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}}.$$

In what follows the upper bound will be of importance: with probability $1 - \eta$,

$$P(\alpha_i) \leq \nu(\alpha_i) + \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}}$$

is valid simultaneously for all decision rules (including those which minimize empirical risk).

§4 A Deterministic Statement of the Problem

The size of the confidence interval computed based on Theorem 6.1 may be excessive. Indeed, consider the case when the set consisting of N decision rules contains a rule which solves perfectly the problem of pattern recognition, i.e., a rule for which the possibility of erroneous classification will equal zero. Such a formulation of the problem is sometimes called *deterministic.*† Then this rule (or a rule close to it) should be found from the sample $x_1, \omega_1; \ldots; x_l, \omega_l$.

We seek this rule using the method of minimizing the empirical risk. Since there exists among functions $F(x, \alpha_i)$ $(i = 1, \ldots, N)$ a function which solves the problem perfectly, it is clear *a priori* that for any sample $x_1, \omega_1; \ldots; x_l, \omega_l$ the value of the minimum of empirical risk will be zero. This minimum, however, can be obtained for several functions. Thus it becomes necessary to estimate the probability that the quality of any function which yields a value of zero for the empirical risk will not be worse than the given $\varkappa$.

Introduce the function

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

† The terminology is unfortunate, since the problem remains statistical. However, we use it because it is widespread.

Then an estimate of the rate of uniform convergence of frequencies to probabilities over the set of events for which the frequency of errors is zero is an estimate of the probability of an event

$$\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)|\,\bar{\theta}(\nu(\alpha_i)) > \varkappa\right\}$$

(rather than the event $\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}$ as in Theorem 6.1).

Since the number of functions for which the zero value of empirical risk is attained does not exceed N (the total number of the functions in this class), the inequality

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)|\bar{\theta}(\nu(\alpha_i)) > \varkappa\right\} \leq NP_\varkappa \tag{6.21}$$

is valid. Here $P_\varkappa$ is the probability that the decision rule for which the probability of committing an error exceeding $\varkappa$ will classify correctly all the elements of the sample. This probability may be easily bounded:

$$P_\varkappa \leq (1 - \varkappa)^l. \tag{6.22}$$

Substituting the bound for $P_\varkappa$ into (6.21), we obtain

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)|\bar{\theta}(\nu(\alpha_i)) > \varkappa\right\} \leq N(1 - \varkappa)^l. \tag{6.23}$$

In order that the probability

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)|\bar{\theta}(\nu(\alpha_i)) > \varkappa\right\}$$

may not exceed the value η, it is sufficient that the equality

$$N(1 - \varkappa)^l = \eta \tag{6.24}$$

be fulfilled. Solving this equation with respect to l, we obtain

$$l = \frac{\ln N - \ln \eta}{-\ln(1 - \varkappa)}. \tag{6.25}$$

Since for small $\varkappa$ the approximation

$$-\ln(1 - \varkappa) \approx \varkappa$$

is valid, (6.25) may be represented in the form

$$l = \frac{\ln N - \ln \eta}{\varkappa}.$$

In contrast with (6.20), the denominator here is $\varkappa$ rather than $2\varkappa^2$, i.e., in the deterministic formulation the sufficient size of the sample is smaller than in the general case. Solving (6.24) with respect to $\varkappa$, we obtain

$$\varkappa = \frac{\ln N - \ln \eta}{l}.$$

Thus the following theorem is valid:

Theorem 6.2. *If one chooses from the set of decision rules consisting of N elements a rule that commits no errors in the sample, then with probability* $1 - \eta$ *one can assert that the probability of erroneous classification using the selected rule is within the limits*

$$0 \leq P \leq \varkappa,$$

where

$$\varkappa = \frac{\ln N - \ln \eta}{l}.$$

§5 Upper Bounds on Error Probabilities

Despite their apparent simplicity, Theorems 6.1 and 6.2 are quite deep. Essentially the subsequent development of the theory of minimizing empirical risk consists of a generalization of these theorems to the case of infinitely many decision rules. The basic points of this further theory are already available. We shall dwell on them in some detail.

(1) Theorems 6.1 and 6.2 are immediately obtained from the bounds on the rate of uniform convergence, over a class of events, of frequencies to probabilities. Theorem 6.1 is based on the bound (6.16) on the rate of uniform convergence over the class of events S_N: $A_1, \ldots, A_N$ of frequencies towards probabilities. Theorem 6.2 is based on a bound on the rate of uniform convergence over a narrower class $\{|P(\alpha_i) - \nu(\alpha_i)|\bar{\theta}(\nu(\alpha_i)) \leq \varkappa\}$. Denote this class by $\hat{S}_N$.

(2) In both cases the rate of uniform convergence was determined by the product of two quantities: the number of events in a class, and a bound on the probability that the frequency of any fixed event in the class deviates by more than $\varkappa$ from the probability of this event. For the events considered in Theorem 6.1 this probability does not exceed $\exp\{-2\varkappa^2 l\}$; for the events considered in Theorem 6.2 the analogous probability does not exceed $(1 - \varkappa)^l \approx \exp\{-\varkappa l\}$. Thus a bound on the rate of uniform convergence of frequencies to probabilities over a class of events is obtained from a bound on the rate of the ordinary convergence which follows from the law of large numbers, by multiplying it by the number of events in this class. When constructing a theory of uniform convergence over a class of events with an infinite number of members, this structure of a bound on the rate of uniform

convergence is retained. However, instead of the number of events, in this case other characteristics of the "capacity" of the class of events are utilized.

(3) In Theorem 6.1 two-sided bounds on the probability of erroneous classification using a decision rule which minimizes the empirical risk were obtained. However, for the subsequent theory the lower bound is of little importance. Therefore it is of interest to obtain a bound on a uniform one-sided deviation, i.e., a bound on

$$P\left\{\sup_i (P(\alpha_i) - \nu(\alpha_i)) > \varkappa\right\},$$

and not on

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\}.$$

The probability of the event $\{\sup_i(P(\alpha_i) - \nu(\alpha_i)) > \varkappa\}$ does not exceed the probability of the event $\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}$. Consequently a more refined bound on the probability of a uniform one-sided deviation $P\{\sup_i (P(\alpha_i) - \nu(\alpha_i)) > \varkappa\}$, than that on the probability of a two-sided uniform deviation $P\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}$ is possible. Such a bound allows us to obtain from the above a bound on the probability of erroneous classification which is better than the one obtained from Theorem 6.1.

(4) The bounds on the rate of uniform convergence given by (6.16) and (6.23) depend substantially on bounds on the probability of deviation of a frequency from the probability of events in the class under consideration (S_N or $\hat{S}_N$). The least favorable event A for the class S_N is that for which $P(A) = \frac{1}{2}$. Therefore only the bound (6.16) is possible. For the class of events $\hat{S}_N$ the least favorable event is the one for which $P(A) = \varkappa$. The more refined bound (6.22) is available for the probability of deviation of the frequency from the probability of this event. Thus the bounds obtained for the classes of events S_N and $\hat{S}_N$ differ in the same manner as the bound on the probability of a deviation of an event A such that $P(A) = \frac{1}{2}$ differs from the corresponding bound on an event A' such that $P(A') = \varkappa$. This fact demands that more careful attention be given to the requirements imposed on the amounts of deviation of frequencies from the respective probabilities for different events in the class. For our purposes of obtaining a uniform bound on the risk it is reasonable not to require a uniform deviation of frequencies from probabilities for all events in the class but to allow a larger deviation for events such that $P(A)$ is close to $\frac{1}{2}$ and a smaller one for events such that $P(A')$ is close to $\varkappa$. For example, it makes sense to bound the uniform relative value of the deviation

$$\left\{\sup_i \frac{P(\alpha_i) - \nu(\alpha_i)}{\sigma(\alpha_i)} > \varkappa\right\},$$

where $\sigma(\alpha_i) = \sqrt{P(\alpha_i)(1 - P(\alpha_i))}$; for small $P(\alpha_i)$ the approximation $\sigma(\alpha_i) \approx \sqrt{P(\alpha_i)}$ is valid. We now obtain a bound on the probability of the one-sided relative deviation

$$P\left\{\sup_i \frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} > \varkappa\right\}, \tag{6.26}$$

and using it we shall construct an upper bound on the probability of erroneous classification. To derive the bound (6.26) we shall utilize the inequality

$$P\left\{\frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} > \varkappa\right\} < \exp\{-\tfrac{1}{2}\varkappa^2 l\}. \tag{6.27}$$

It follows from (6.27) that for a class consisting of N events the following bound on the rate of uniform convergence is valid:

$$P\left\{\sup_i \frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} > \varkappa\right\} < N\exp\{-\tfrac{1}{2}\varkappa^2 l\}. \tag{6.28}$$

We shall require that the probability of uniform one-sided relative deviation (6.28) not exceed η:

$$N\exp\{-\tfrac{1}{2}\varkappa^2 l\} = \eta.$$

This is certainly satisfied if

$$\varkappa = \sqrt{2\frac{\ln N - \ln \eta}{l}}. \tag{6.29}$$

Let the condition (6.29) be fulfilled. Then the inequality

$$\frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} < \varkappa \tag{6.30}$$

is satisfied simultaneously for all events A_i with probability $1 - \eta$. Solving (6.30) for $P(\alpha_i)$, we obtain that

$$P(\alpha_i) < \frac{\varkappa^2}{2}\left(1 + \sqrt{1 + \frac{4\nu(\alpha_i)}{\varkappa^2}}\right) + \nu(\alpha_i) \tag{6.31}$$

is valid with probability $1 - \eta$ for all the events in the class simultaneously.

Substituting (6.29) into (6.31), we obtain that with probability $1 - \eta$, the N simultaneous inequalities

$$P(\alpha_i) \leq \frac{\ln N - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2\nu(\alpha_i)l}{\ln N - \ln \eta}}\right) + \nu(\alpha_i)$$

are fulfilled. We have thus proved the following theorem:

Theorem 6.3. *Let the set of decision rules consist of N elements, and for each rule $F(x, \alpha_i)$ let the frequency of errors in the sample equal $\nu(\alpha_i)$. Then one can* *assert with probability* $1 - \eta$ *that the bounds*

$$P(\alpha_i) \leq \frac{\ln N - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2\nu(\alpha_i)l}{\ln N - \ln \eta}}\right) + \nu(\alpha_i) \qquad (6.32)$$

are fulfilled simultaneously for all decision rules in the class.

Remark. Since the bound (6.32) is valid, with probability $1 - \eta$, simultaneously for all the rules in the class, it also holds for the rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk.

Theorem 6.3 allows us to estimate the quality of the rule which minimizes the empirical risk. Moreover, the bound (6.32) coincides with the bound given in Theorem 6.2 obtained in the extreme case when $P(\alpha^*) \approx 0$, and it is close to the bound given in Theorem 6.1 for the second extreme case when $P(\alpha^*) \approx \frac{1}{2}$. The structure of bounds for an infinite class of decision rules is the same.

§6 An ε-net of a Set

In the preceding sections we established the existence of a uniform convergence of frequencies of occurrences of events to the corresponding probabilities over a class of events consisting of a finite number of elements; we obtained bounds on the rate of this convergence and using it, bounds on the quality of a decision rule which minimizes the empirical risk. Our task is to generalize these results to the case of infinitely many events.

In general, however, in the infinite case the uniform convergence of frequencies to probabilities may not occur: for example, if the set of events is defined as consisting of all open subsets of the set X, ω. In this case a situation may arise where (cf. the example in Section 2) an algorithm for minimizing the empirical risk yields the value zero for the risk but it is not capable of learning. Therefore the problem is to determine conditions which will assure uniform convergence for an infinite number of events, to bound its rate, and finally to obtain an upper bound on the probability of erroneous classification for a rule which minimizes the empirical risk.

In mathematics the necessity often arises of extending results valid for a finite set of elements to the infinite case. Usually such a generalization is possible if the infinite set can be covered by a *finite ε-net*.

Definition. The set B of elements in a metric space R is called an *ε-net* of the set G if any point $c \in G$ is distant from some point $b \in B$ by an amount not exceeding ε, i.e., $\rho(b, c) < \varepsilon$.

We say that the set G *admits a covering by a finite ε-net* if for each ε there exists an ε-net B consisting of a finite number of elements.

In this section, for an infinite set of decision rules admitting a covering by a finite ε-net we shall obtain assertions analogous to the assertions of Theorems 6.1 and 6.3.

Thus let an infinite set of decision rules $F(x, \alpha)$ be given on which the metric $\rho(\alpha_1, \alpha_2) = \rho(F(x, \alpha_1), F(x, \alpha_2))$ is defined and a finite ε-net is singled out. Let this finite ε-net consist of $N(\varepsilon)$ elements. Moreover, let it be given that if two decision rules $F(x, \alpha_1)$ and $F(x, \alpha_2)$ are distant from each other by an amount not exceeding ε ($\rho(\alpha_1, \alpha_2) \leq \varepsilon$), then the quality of these rules differs by an amount not exceeding $\delta(\varepsilon)$, i.e.,

$$\left|\int (\omega - F(x, \alpha_1))^2 P(x, \omega)\, dx\, d\omega - \int (\omega - F(x, \alpha_2))^2 P(x, \omega)\, dx\, d\omega\right| \leq \delta(\varepsilon).$$

In other words, a small variation in the decision rule implies a small variation in the quality of classification.

Under these conditions Theorems 6.1 and 6.3 can be generalized as follows:

Theorem 6.4. *Let the set of decision rules $F(x, \alpha)$ be covered by a finite ε-net. Then with probability $1 - \eta$ the quality of the decision rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk is bounded by*

$$\nu(\alpha_i(\alpha_{\text{emp}})) - \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} - \delta(\varepsilon) \leq P(\alpha_{\text{emp}})$$

$$\leq \nu(\alpha_i(\alpha_{\text{emp}})) + \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} + \delta(\varepsilon),$$

where $F(x, \alpha_i(\alpha_{\text{emp}}))$ is an element of the ε-net which is closest to $F(x, \alpha_{\text{emp}})$.

Theorem 6.5. *Let the set of decision rules $F(x, \alpha)$ be covered by a finite ε-net. Then with probability $1 - \eta$ the quality of the decision rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk is bounded by*

$$P(\alpha_{\text{emp}}) \leq \nu(\alpha_i(\alpha_{\text{emp}})) + \frac{\ln N(\varepsilon) - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2\nu(\alpha_i(\alpha_{\text{emp}}))l}{\ln N(\varepsilon) - \ln \eta}}\right) + \delta(\varepsilon),$$

where $F(x, \alpha_i(\alpha_{\text{emp}}))$ is an element of the ε-net which is closest to $F(x, \alpha_{\text{emp}})$.

Remark. Theorems 6.4 and 6.5 are valid for any ε-net given *a priori* (before the appearance of the sample). In particular the value of ε which defines the ε-net can be chosen in Theorem 6.4 from the condition of the minimum of expression

$$\sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} + \delta(\varepsilon),$$

and in Theorem 6.5 from the condition of the minimum of expression

$$\frac{\ln N(\varepsilon) - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2cl}{\ln N(\varepsilon) + \ln \eta}}\right) + \delta(\varepsilon),$$

where $0 \leq c \leq 1$ is a constant (for example $c = 0.5$).

Theorems 6.4 and 6.5 are proved in the same way:

Proof.

(1) A finite ε-net consisting of $N(\varepsilon)$ elements

$$F(x, \alpha_1), \ldots, F(x, \alpha_{N(\varepsilon)}) \tag{6.33}$$

is given for the set of decision rules $F(x, \alpha)$. According to Theorem 6.1 (6.3) the inequalities

$$\begin{gathered} \nu(\alpha_i) - \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} \le P(\alpha_i) \le \nu(\alpha_i) + \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}}, \\ \left(P(\alpha_i) \le \frac{\ln N(\varepsilon) - \ln \eta}{l} \left(1 + \sqrt{1 + \frac{2\nu(\alpha_i) l}{\ln N(\varepsilon) - \ln \eta}} \right) + \nu(\alpha_i) \right) \end{gathered} \tag{6.34}$$

are fulfilled with probability $1 - \eta$ simultaneously for all $N(\varepsilon)$ elements of (6.33).

(2) For any decision rule $F(x, \alpha^*)$ (including the one which minimizes in $F(x, \alpha)$ the value of the empirical risk), the closest element of the ε-net $F(x, \alpha_i(\alpha^*))$ can be found, for which this element satisfies

$$|P(\alpha^*) - P(\alpha_i(\alpha^*))| \le \delta(\varepsilon). \tag{6.35}$$

The inequalities (6.34) and (6.35) imply that for the decision rule $F(x, \alpha_i(\alpha_{\text{emp}}))$ the relations

$$\begin{gathered} \nu(\alpha_i(\alpha_{\text{emp}})) - \sqrt{\frac{\ln N - \ln(\eta/2)}{2l}} - \delta(\varepsilon) \\ \le P(\alpha_{\text{emp}}) \le \nu(\alpha_i(\alpha_{\text{emp}})) + \sqrt{\frac{\ln N(\varepsilon) - \ln(\eta/2)}{2l}} + \delta(\varepsilon), \\ \left(P(\alpha_{\text{emp}}) \le \frac{\ln N(\varepsilon) - \ln \eta}{l} \left(1 + \sqrt{1 + \frac{2\nu(\alpha_i(\alpha_{\text{emp}}))l}{N(\varepsilon) - \ln \eta}} \right) + \delta(\varepsilon) + \nu(\alpha_i(\alpha_{\text{emp}})) \right) \end{gathered}$$

are valid with probability $1 - \eta$. The theorems are thus proved. □

Thus if the set of decision rules $F(x, \alpha)$ admits a cover by a finite ε-net and the distribution $P(x, \omega)$ is such that close values of the probability of erroneous classification correspond to close decision rules, then as the sample size increases the method of minimizing the empirical risk should in principle successfully yield the desired result.† Moreover for each fixed ε the probability of erroneous classification using the rule which minimizes the empirical risk is bounded in terms of the inequalities (6.34).

However, in order to utilize these bounds the value of $\delta(\varepsilon)$ is required. To compute this value the density $P(x)$ is used, which in the formulation of the problem of pattern recognition is assumed to be unknown. In the next chapter, when solving the problem of estimating regression, we shall obtain the value of $\delta(\varepsilon)$ and be able to utilize bounds on the quality of a function expressed in terms of the value of empirical risk $\delta(\varepsilon)$ and $N(\varepsilon)$. In this chapter, to obtain the rate of uniform convergence of frequencies to the respective probabilities over an infinite class of events, a new idea will be utilized. This will eventually lead us to the construction of necessary and sufficient conditions for uniform convergence, to the derivation of a bound on the rate of uniform convergence based on these conditions, and finally to a constructive bound on the quality of a decision rule obtained using the method of minimizing the empirical risk.

† Although this assertion does not follow formally from Theorem 6.4, its proof is completely analogous.

§7 Necessary and Sufficient Conditions for Uniform Convergence of Frequencies to Probabilities

Up until now we have utilized quite rough "capacity" characteristics of the set of decision rules (the number of elements in the set) to obtain bounds on the rate of uniform convergence. In this section we introduce a more refined characteristic of capacity—*the entropy of a system of events on samples of size l.* Using this characteristic one can establish exhaustive necessary and sufficient conditions for uniform convergence of frequencies of events to their respective probabilities, i.e., for the equality

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

to be valid for any $\varkappa$.

Thus let a set S of decision rules $F(x, \alpha)$ be defined and a sample $x_1, \ldots, x_l$ be given. This sample can generally be subdivided into two classes in 2^l ways. However, only those subdivisions of the sample which can be accomplished using the rules $F(x, \alpha)$ will be of interest. (Uusing the rule $F(x, \alpha^*)$, the set $x_1, \ldots, x_l$ is subdivided into two subsets: one on which $F(x, \alpha^*) = 1$, and the other on which $F(x, \alpha^*) = 0$.) The number of different subdividing methods depends on the class of decision rules $F(x, \alpha)$ as well as on the sample. We shall denote this number by

$$\Delta^S(x_1, \ldots, x_l).$$

Consider the system of events

$$S(\alpha) = \{x, \omega : (\omega - F(x, \alpha))^2 = 1\}$$

formed by the set of decision rules $F(x, \alpha)$. Let a random independent sample

$$x_1, \omega_1; \ldots; x_l, \omega_l \tag{6.36}$$

be given. The system of events $S(\alpha)$ induces $\Delta(S(\alpha); x_1, \omega_1; \ldots; x_l, \omega_l)$ different subsamples on the sample (6.36). Clearly the number of these subsamples equals $\Delta^S(x_1, \ldots, x_l)$. Since $x_1, \ldots, x_l$ is a random independent sample the number of subdivisions $\Delta^S(x_1, \ldots, x_l)$ is a random variable.

Definition. The quantity

$$H^S(l) = M \ln \Delta^S(x_1, \ldots, x_l)$$

is called the *entropy* of a system of events $S(\alpha)$ on a sample of size l.

It turns out that for the uniform convergence of frequencies $\nu(\alpha)$ to the respective probabilities $P(\alpha)$ over the set of events, it is necessary and sufficient that as the sample size increases, the portion of the entropy due to a single element of the sample approach zero, i.e., that the sequence

$$\frac{H^S(1)}{1}, \frac{H^S(2)}{2}, \ldots, \frac{H^S(l)}{l}$$

approach zero as l increases. In other words the condition

$$\lim_{l \to \infty} \frac{H^S(l)}{l} = 0 \tag{6.37}$$

should be fulfilled. The proof of this assertion follows from Theorem A.1 of the Appendix to Chapter 7.

Like any exhaustive conditions, the necessary and sufficient conditions stated above for the uniform convergence of frequencies to their respective probabilities utilize some refined concepts. In our case such a concept is the entropy $H^S(l)$ of a system of events $S(\alpha)$ on samples of size l, which is constructed by means of the density $P(x)$. In the case of the problem of pattern recognition the density is unknown, as stated above. Therefore, in order to establish the feasibility of minimizing the expected risk via the determination of the minimum of empirical risk, the necessary and sufficient conditions (6.37) cannot be used.

For this reason it is important to obtain less refined sufficient conditions which firstly will not depend on the properties of the measure $P(x)$ and secondly will admit a bound on the rate of uniform convergence. Such conditions may be stated in terms of a capacity measure of the system of events $S(\alpha)$ which is obtained from the entropy $H^S(l)$ by abstracting it from measure properties.

Definition. The function

$$m^S(l) = \max_{x_1, \ldots, x_l} \Delta^S(x_1, \ldots, x_l),$$

where the maximum is taken over all possible samples of size l, is called the *growth function* of a system of events formed by the decision rules $F(x, \alpha)$.

The growth function is constructed in such a manner that it does not depend on the properties of measure $P(x)$ and the inequality

$$\ln m^S(l) \geq H^S(l) \tag{6.38}$$

is always satisfied. Now if the quantity

$$\frac{\ln m^S(l)}{l}$$

approaches zero as l increases, then in view of (6.38) the ratio $H^S(l)/l$ tends to zero *a fortiori*. Therefore the condition

$$\lim_{l \to \infty} \frac{\ln m^S(l)}{l} = 0$$

is a sufficient condition for the uniform convergence of frequencies to their probabilities. Below we shall show that the growth function can be easily obtained for the events defined by various classes of decision rules $F(x, \alpha)$ and hence the uniform convergence can be established. Moreover, as will be shown below, the rate of uniform convergence can also be estimated using the growth function $m^S(l)$.

§8 Properties of Growth Functions

A growth function has a simple interpretation: it counts the maximal number of ways for subdividing l points into two classes using the decision rules $F(x, \alpha)$. For growth functions the following remarkable theorem is valid.

Theorem 6.6. *A growth function is either identically equal to 2^l or for $l > h$ is majorized by the function*

$$m^S(l) < 1.5 \frac{l^h}{h!},$$

where $h + 1$ is the smallest sample size such that the condition $m^S(l) = 2^l$ is violated. In other words

$$m^S(l) \begin{cases} \text{either} \equiv 2^l, & \\ \text{or} \quad < 1.5 \frac{l^h}{h!} & (l > h). \end{cases}$$

The proof of this theorem is presented in the appendix to this chapter.

In order to bound a growth function it is necessary to show that either (1) for any l, points $x_1, \dots, x_l$ exist such that using the decision rules $F(x, \alpha)$ it would be possible to subdivide them into two classes by any one of the 2^l ways, or (2) a number h exists such that h points can be subdivided into classes in all possible ways, but $h + 1$ points cannot. In the first case the growth function is exponential; in the second it is polynomial. The number h can serve as the measure of diversity of the class of decision rules.

Definition. We say that the class of indicator functions *has capacity h* if the inequality

$$m^S(l) < 1.5 \frac{l^h}{h!} \qquad (l > h) \tag{6.39}$$

is valid. If the equality

$$m^S(l) \equiv 2^l$$

is satisfied we say that the capacity h of the class of indicator functions $F(x, \alpha)$ is *infinite*.

It is easy to verify that if the capacity of the class of indicator functions is finite, then the uniform convergence of frequencies to the respective probabilities always occurs. Indeed, in this case the relation

$$0 \leq \lim_{l \to \infty} \frac{\ln m^S(l)}{l} \leq \lim_{l \to \infty} \frac{h \ln l - \sum_{i=1}^{h} \ln i}{l} = 0$$

is valid and the sufficient condition is fulfilled.

The following class of decision rules, which are linear in the parameter, plays an important role in the subsequent theory:

$$F(x, \alpha) = \theta\left(\sum_{i=1}^{n} \alpha_i \varphi_i(x)\right); \qquad \theta(z) = \begin{cases} 1 & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases} \tag{6.40}$$

It is easy to obtain a growth function for a class of events defined by linear decision rules (6.40). For this purpose it is sufficient to determine the maximal number h of points in the space of dimensionality n which can be subdivided into two classes using a hyperplane in any one of the 2^h ways. It is known that this number equals n. Therefore according to Theorem 6.6 the growth function is bounded by

$$m^S(l) < 1.5 \frac{l^n}{n!} \qquad (l > n)$$

for the class of linear decision rules (6.40). Consequently for the class of linear decision rules sufficient conditions for uniform convergence are fulfilled.

It was shown in Chapter 2 that uniform convergence of frequencies of events to their probabilities over a class of events defined by one-dimensional linear decision rules $F(x, \alpha) = \theta(x + \alpha)$ makes up the content of the Glivenko–Cantelli theorem, which asserts the uniform convergence of the empirical cumulative distribution function to the population one.

§9 Bounds on Deviations of Empirically Optimal Decision Rules

In the appendix to this chapter a bound on the rate of uniform convergence of frequencies to probabilities over a class of events $S(\alpha)$ is obtained. It is shown that the inequality

$$P\left\{\sup_{\alpha} |P(\alpha) - v(\alpha)| > \varkappa\right\} < 6m^S(2l) \exp\left\{-\frac{\varkappa^2 l}{4}\right\} \tag{6.41}$$

is valid. The bound (6.41) is of the same form as the above: it is formed by multiplying the quantity $6m^S(2l)$—which is the capacity characteristic of the system of events—by a bound on the probability that the deviation of the frequency from its probability exceeds $\varkappa$ (the quantity $\exp\{-\varkappa^2 l/4\}$).

If the capacity of the class of decision rules is infinite ($m^S(l) \equiv 2l$), then the bound (6.41) is trivial, since for all $\varkappa$ the right-hand side of the inequality exceeds 1. The bound (6.41) is meaningful when the capacity of the class of decision rules is finite:

$$m^S(l) < 1.5 \frac{l^h}{h!}.$$

In this case it takes the form

$$P\left\{\sup_\alpha |P(\alpha) - v(\alpha)| > \varkappa\right\} < 9\frac{(2l)^h}{h!}\exp\left\{-\frac{\varkappa^2 l}{4}\right\}. \tag{6.42}$$

As l increases, the right-hand side of the inequality (6.42) tends to zero and the approach is faster for smaller values of the capacity h. We shall require that the probability

$$P\left\{\sup_\alpha |P(\alpha) - v(\alpha)| > \varkappa\right\}$$

not exceed η. This is certainly true if

$$9\frac{(2l)^h}{h!}\exp\left\{-\frac{\varkappa^2 l}{4}\right\} = \eta. \tag{6.43}$$

Equation (6.43) can be solved for κ (using Stirling's formula):

$$\varkappa = 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}}. \tag{6.44}$$

Then (6.42)–(6.44) imply the following theorem:

Theorem 6.7. *Let $F(x, \alpha)$ be the class of decision rules of bounded capacity h, and let $v(\alpha)$ be the frequency of errors computed from the sample for the rule $F(x, \alpha)$. Then with probability $1 - \eta$ one may assert that for $l > h$, and simultaneously for all decision rules $F(x, \alpha)$, the probability of erroneous classification is within the limits*

$$v(\alpha) - 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}}$$

$$< P(\alpha) < v(\alpha) + 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}}.$$

Remark. Theorem 6.7 implies that for the rule $F(x, \alpha_{\text{emp}})$, which minimizes the empirical risk, the upper bound

$$P(\alpha_{\text{emp}}) < \nu(\alpha_{\text{emp}}) + 2\sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{9}}{l}} \qquad (l > h)$$

is valid with probability $1 - \eta$.

In the appendix to this chapter it is shown that along with (6.41) the bound

$$P\left\{\sup_{\alpha} \frac{P(\alpha) - \nu(\alpha)}{\sqrt{P(\alpha)}} > \varkappa\right\} < 8m^s(2l)e^{-\varkappa^2 l/4}$$

is valid. This bound is nontrivial for a class of decision rules of bounded capacity:

$$P\left\{\sup_{\alpha} \frac{P(\alpha) - \nu(\alpha)}{\sqrt{P(\alpha)}} > \varkappa\right\} < 12\frac{(2l)^h}{h!}e^{-\varkappa^2 l/4}. \tag{6.45}$$

We shall require that the right-hand side of the inequality be equal to η:

$$12\frac{(2l)^h}{h!}e^{-\varkappa^2 l/4} = \eta.$$

This is fulfilled if

$$\varkappa = 2\sqrt{\frac{\ln \frac{(2l)^h}{h!} - \ln \frac{\eta}{12}}{l}} \approx 2\sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{12}}{l}}. \tag{6.46}$$

On the other hand, the inequality (6.45) can be stated as follows: with probability η, simultaneously for all α the inequality

$$P(\alpha) \le \frac{\varkappa^2}{2}\left(1 + \sqrt{1 + \frac{4\nu(\alpha)}{\varkappa^2}}\right) + \nu(\alpha) \tag{6.47}$$

is valid. The relations (6.46) and (6.47) imply the following theorem.

Theorem 6.8. *Let $F(x, \alpha)$ be a class of decision rules of bounded capacity h, and for each rule $F(x, \alpha)$ let the frequency of errors computed in the sample equal $\nu(\alpha)$. Then with probability $1 - \eta$ one can assert that the bound*

$$P(\alpha) \le 2\,\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{12}}{l}\left(1 + \sqrt{1 + \frac{\nu(\alpha)l}{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{12}}}\right) + \nu(\alpha_{\text{emp}}) \tag{6.48}$$

is valid for $l > h$ simultaneously for all rules in the class.

Remark. It follows from Theorem 6.8 that for the rule $F(x, \alpha_{\text{emp}})$ which minimizes the empirical risk the bound

$$P(\alpha_{\text{emp}}) \leq 2\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{12}}{l}\left(1+\sqrt{1+\frac{\nu(\alpha_{\text{emp}})l}{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{12}}}\right)+\nu(\alpha_{\text{emp}})$$

is valid.

§10 Remarks on the Bound on the Rate of Uniform Convergence of Frequencies to Probabilities

In this chapter we have obtained bounds on the rate of uniform convergence of frequencies to the respective probabilities:

$$P\left\{\sup_{\alpha}|P(\alpha)-\nu(\alpha)|>\varkappa\right\}<\begin{cases}2Ne^{-2\varkappa^2 l},\\ 6m^s(2l)e^{-\varkappa^2 l/4},\end{cases}$$

and bounds on the uniform one-sided relative deviations of frequencies from their probabilities:

$$P\left\{\sup_{\alpha}\frac{P(\alpha)-\nu(\alpha)}{\sqrt{P(\alpha)}}>\varkappa\right\}<\begin{cases}Ne^{-\varkappa^2 l/2},\\ 8m^s(2l)e^{-\varkappa^2 l/4}\end{cases}$$

Using these bounds, Theorems 6.1, 6.3, 6.7, and 6.8 were obtained, which allow us to estimate the quality of a decision rule minimizing the empirical risk.

All the estimates obtained have the same structure, consisting of two factors: one which bounds the probability of the corresponding deviation (separately) for each event in the class, and another which characterizes the variety of the class of decision rules. Different characteristics of the variety of the class of decision rules are used for the bounds. The simplest is the number of decision rules in the class. The simplicity of this characteristic is due to the fact that it does not, for example, take into account whether the decision rules in the class are "substantially different" or whether all the rules are "equivalent."

An adequate measure of the variety of the class of decision rules, by which it is possible to construct necessary and sufficient conditions for the uniform convergence of frequencies to their probabilities, is the entropy of the system of events defined by the decision rules. However, to compute the entropy of a system of events on samples of length l is possible only if the density $P(x)$ is known, and it is assumed to be unknown in the formulation of the pattern recognition problem. Therefore a new measure of variety was introduced which is obtained from entropy by choosing the least favorable distribution. This measure is expressed in terms of the capacity of the class of decision rules and can easily be computed.

Various definitions of measures of variety of a class of decision rules generate different theorems on the quality of algorithms minimizing the empirical risk. However, in all these theorems the very same fact is asserted: if the measure of variety of a class of decision rules is small compared with the sample size, then the method of minimizing empirical risk allows us to choose a rule which is close to the best one in the class.

A characteristic feature of the theory of minimizing empirical risk presented above is the complete absence of any indications as to the constructive feasibility of determining an algorithm. This feature has negative as well as positive aspects. On one hand, the theory does not give regular procedures for minimizing empirical riks; they should be implemented by a corresponding program. On the other hand, the theory is quite general. The method can be applied to various classes of decision rules: linear discriminant functions, piecewise linear discriminant functions, logistic functions of a particular kind, and so on. This is due to the fact that the theory of the method of minimizing empirical risk answers the question "what to do," leaving the question "how to do it" unsettled. Therefore various methods can be applied, including heuristic ones.

The application of heuristic methods in this case has some theoretical justification: if in a class of decision rules whose capacity is small compared to the sample size one chooses a rule which, while it does not yield the minimum of the empirical risk, results in a sufficiently small value of it, then in view of the theorems proved above, the decision rule selected will be of sufficiently high quality.

Constructive ideas for such algorithms admit a simple geometric interpretation: It is required to construct in a space X a hypersurface belonging to a given class of hypersurfaces which—with the smallest possible number of errors—will separate the vectors of the sample in one class from the corresponding vectors in the other. The assignment of vectors (including those which do not belong to a learning sequence) to a particular class is carried out according to the side of the subdividing hypersurface on which the vector is located.

Methods of constructing separating hypersurfaces constitute a constructive part of the theory of pattern recognition. These methods are presented in Addendum I.

§11 Remark on the General Theory of Uniform Estimating of Probabilities

We have thus developed a theory of uniform estimating of error probabilities in pattern recognition for arbitrary classes of decision rules. Formally, in the functional which computes the probabilities of errors we wrote a quadratic loss function. In proving the related theorems, however, the form of the loss function was unimportant. What is important is that $Q(z, \alpha)$, $\alpha \in \Lambda$, is a class of indicator functions.

In fact, this chapter presents a theory more general than the uniform estimation of error probabilities in pattern recognition. Here a general theory has been developed for uniform estimation of probabilities from their frequencies in a class of events of limited capacity. We now formulate the basic assertions of this theory. The proofs are identical to those of similar theorems given in the chapter.

Assume that a space Z is given on which a probability measure $P(z)$ has been defined and a system of events S_α, $\alpha \in \Lambda$, is specified (subsets measurable with respect to the given measure and belonging to Z). Let $Q(z, \alpha)$, $\alpha \in \Lambda$ be a family of indicator functions on the sets S_α, $\alpha \in \Lambda$ (i.e., the function

$$Q(z, \alpha) = \left\{\begin{matrix} 0 & \text{if } z \notin S_\alpha \\ 1 & \text{if } z \in S_\alpha \end{matrix}\right).$$

Let the capacity of the family of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$, be finite and equal to h (there exists such an h that $m^{S_\alpha}(h) = 2^h$, $m^{S_\alpha}(h + 1) \neq 2^{h+1}$).

Under these conditions the following assertions hold on two-sided and one-sided uniform bounds of probabilities

$$P(\alpha) = \int_{S_\alpha} dP(z) = \int Q(z, \alpha)\, dP(z)$$

by virtue of associated frequencies

$$\nu(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha)$$

computed on a sample

$$z_1, \ldots, z_l.$$

Assertion 1. *For any $l > (\Delta/(\Delta - 1))^2$, $\Delta > 1$ with probability $1 - \eta$ simultaneously for all events S_α, $\alpha \in \Lambda$, the two-sided bound*

$$\nu(\alpha) - \Delta\sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{9}}{l}} \leq P(\alpha) \leq \nu(\alpha) + \Delta\sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{9}}{l}}$$

holds.

Assertion 2. *With probability* $1 - \eta$ *simultaneously for all events* S_α, $\alpha \in \Lambda$, *the one-sided bound*

$$P(\alpha) \leq \nu(\alpha) + 2\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{12}}{l}\left(1 + \sqrt{1 + \frac{\nu(\alpha)l}{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{12}}}\right)$$

holds.

Appendix to Chapter 6

Theory of Uniform Convergence of Frequencies to Probabilities: Sufficient Conditions†

§A1 Sufficient Conditions for Uniform Convergence of Frequencies to Probabilities

According to Bernoulli's classical theorem the frequency of occurrence of a certain event A in a sequence of independent trials converges (in probability) to the probability of this event. Often, however, it becomes necessary to assess simultaneously the probabilities of a class of events S based on the very same sample. Moreover, it is required that the frequencies converge to the probabilities uniformly over all events in the class S. More precisely, the probability that the maximal deviation over the class of frequencies from probabilities exceeds a given, arbitrarily small positive constant must tend to zero as the number of trials increases indefinitely.

It turns out that even in the simplest cases uniform convergence may not occur. Therefore a criterion is required which will test whether such convergence is present.

Let X be a set of elementary events on which a probability measure $P(x)$ is defined. Let S be a collection of random events, i.e., subsets of a space measurable with respect to the measure $P(x)$ (S is included in the σ-algebra of random events, but does not necessarily coincide with it). Denote by $X(l)$ the space of random independent samples taken from X of length l.

For each sample $X^l = x_1, \ldots, x_l$ and event $A \in S$, the frequency of occurrence of event A is defined as the ratio of the number $n(A)$ of elements of the sample belonging to A to the common sample size l:

$$v^l(A) = v(x_1, \ldots, x_l) = \frac{n(A)}{l}.$$

† Necessary and sufficient conditions for uniform convergence of frequencies to probabilities will follow from the results presented in the Appendix to Chapter 7.

Bernoulli's theorem asserts that for a fixed event A the deviation of the frequency from the probability tends to zero (in probability) with increasing sample size, i.e., for any $\varkappa$

$$P\{|P(A) - \nu^l(A)| > \varkappa\} \underset{l\to\infty}{\longrightarrow} 0.$$

Here, however, we are concerned with the maximal (over the class S) deviation of the frequency from the probability:

$$\pi(l) = \sup_{A\in S} |\nu^l(A) - P(A)|.$$

The quantity $\pi(l)$ is a function of a point in the space $X(l)$. We shall assume that this function is measurable with respect to a measure in $X(l)$, i.e., $\pi(l)$ is a random variable. The theorems below deal with bounds on the probabilities of the event $\pi(l)$.

§A2 The Growth Function

Let X be a set, S be a system of its subsets, and $X^l = x_1, \ldots, x_l$ be a sequence of elements x of length l. Each set $A \in S$ determines a subsequence X_A of this sequence consisting of elements belonging to A. We say that A induces a subsequence X_A on the sequence X^l.

Denote by

$$\Delta^S(x_1, \ldots, x_l)$$

the number of different subsequences X_A induced by the sets $A \in S$. Clearly,

$$\Delta^S(x_1, \ldots, x_l) \le 2^l.$$

The number $\Delta^S(x_1, \ldots, x_l)$ is called the *index of the system S relative to the sample* $x_1, \ldots, x_l$.

The index of a system may be defined in another way as well. We shall consider $A_1 \in S$ to be equivalent to $A_2 \in S$ relative to the sample $x_1, \ldots, x_l$ if $X_{A_1} = X_{A_2}$. Then the index $\Delta^S(x_1, \ldots, x_l)$ is the number of equivalence classes into which the system S is subdivided by this equivalence relation.

Clearly the two definitions are equivalent. The function

$$m^S(l) = \max_{x_1, \ldots, x_l} \Delta^S(x_1, \ldots, x_l), \tag{A.1}$$

where the maximum is taken over all the sequences of length l is called *the growth function of the system S*. Here the maximum is always attained, since the index $\Delta^S(x_1, \ldots, x_l)$ takes on a finite number of values.

The growth function of a class of events possesses the following remarkable property.

Theorem A.1. *The growth function either is identically equal to 2^l or is bounded by the function*

$$\sum_{i=0}^{n-1} C_l^i$$

where n is the minimal value of l such that

$$m^S(l) \neq 2^l.$$

In other words

$$m^S(l) \begin{cases} \text{either} & \equiv 2^l, \\ \text{or} & < \sum_{i=0}^{n-1} C_l^i \end{cases} \tag{A.2}$$

To prove this assertion the following lemma is required.

Lemma A.1. *If for some sequence $x_1, \ldots, x_l$ and some n*

$$\Delta^S(x_1, \ldots, x_l) > \sum_{i=0}^{n-1} C_l^i,$$

then there exists a subsequence X^n of length n such that

$$\Delta^S(X^n) = 2^n.$$

Proof. Denote

$$\sum_{i=0}^{n-1} C_l^i = \Phi(n, l)$$

(here and below we shall assume that $C_l^i = 0$ for $i > l$). For this function, as it is easy to verify, the relations

$$\begin{aligned} &\Phi(1, l) = 1, \\ &\Phi(n, l) = 2^l \quad \text{if } l \leq n + 1, \\ &\Phi(n, l) = \Phi(n, l - 1) + \Phi(n - 1, l - 1), \quad \text{if } n \geq 2, l \geq 1 \end{aligned} \tag{A.3}$$

are valid. In turn these relations uniquely determine the function $\Phi(n, l)$ for $l > 0$ and $n > 0$.

We shall prove the lemma by an induction on l and n. For $n = 1$ and any $l \geq 1$ the assertion of the lemma is obvious. Indeed, in this case

$$\Delta^S(x_1, \ldots, x_l) > 1$$

implies that an element of the sequence x_i exists such that for some $A^* \in S$ we have $x_i \in A^*$, while for some other $A^{**} \in S$ we have $x_i \notin A^{**}$. Consequently,

$$\Delta^S(x_i) = 2.$$

For $l < n$ the assertion of the lemma is valid because the premise is false. Indeed, in this case the premise is

$$\Delta^S(x_1, \dots, x_l) > 2^l,$$

which is impossible, since

$$\Delta^S(x_1, \dots, x_l) \leq 2^l.$$

Finally assume that the lemma is valid for $n \leq n_0$ $(n_0 \geq 1)$ for all l. Consider now the case $n = n_0 + 1$. We show that the lemma is valid in this case also for all l.

We fix $n = n_0 + 1$ and carry out the induction on l. As was pointed out, for $l < l_0 + 1$ the lemma is valid. We shall assume that it is valid for $l \leq l_0$ and show that it is valid for $l = l_0 + 1$. Indeed, let the condition of the lemma,

$$\Delta^S(x_1, \dots, x_{l_0}, x_{l_0+1}) > \Phi(n_0 + 1, l_0 + 1)$$

be fulfilled for some sequence $x_1, \dots, x_{l_0}, x_{l_0+1}$. The lemma will be proved if we find a subsequence of length $n_0 + 1$, say $X^{n_0+1} = x_1, \dots, x_{n_0+1}$, such that

$$\Delta^S(x_1, \dots, x_{n_0+1}) = 2^{n_0+1}.$$

Consider the subsequence $X^{l_0} = x_1, \dots, x_{l_0}$. Two cases are possible:

(a) $\Delta^S(x_1, \dots, x_{l_0}) > \Phi(n_0 + 1, l_0)$,
(b) $\Delta^S(x_1, \dots, x_{l_0}) \leq \Phi(n_0 + 1, l_0)$.

In case (a), in view of the induction assumption, there exists a subsequence of length $n_0 + 1$ such that $\Delta^S(X^{n_0+1}) = 2^{n_0+1}$, q.e.d.

In case (b) we subdivide subsequences of the sequence X^{l_0} induced by the sets in S into two types. We assign to the first type subsequences X^r such that on the whole sequence X^{l_0+1} events belonging to S induce X^r as well as (X^r, x_{l_0+1}). Sequences X^r such that either X^r or (X^r, x_{l_0+1}) is induced on the sequence X^{l_0+1} are assigned to the second type. Denote the number of subsequences of the first type by K_1 and of the second by K_2. It is easy to see that

$$\Delta^S(x_1, \dots, x_{l_0}) = K_1 + K_2,$$

$$\Delta^S(x_1, \dots, x_{l_0}, x_{l_0+1}) = 2K_1 + K_2;$$

and hence

$$\Delta^S(x_1, \dots, x_{l_0}, x_{l_0+1}) = \Delta^S(x_1, \dots, x_{l_0}) + K_1. \tag{A.4}$$

Denote by S' the system of all subsets $A \in S$ that induce subsequences of the first type on the sequence X^{l_0}. Then if

(b′) $K_1 = \Delta^{S'}(x_1, \dots, x_{l_0}) > \Phi(n_0, l_0)$,

in view of the induction assumption there exists a subsequence $X^{n_0} = x_{i_1}, \dots, x_{i_{n_0}}$ such that

$$\Delta^{S'}(x_{i_1}, \dots, x_{i_{n_0}}) = 2^{n_0} \qquad (X^{n_0} \subset X^{l_0}).$$

However, in that case we have

$$\Delta^{S'}(x_1, \ldots, x_{i_{n_0}}, x_{l_0+1}) = 2^{n_0+1}$$

for the sequence $x_{i_1}, \ldots, x_{i_{n_0}}, x_{l_0+1}$, since for each subsequence X^r induced on the sequence X^{n_0}, two subsequences induced on X^r, x_{l_0+1} can be found, namely X^r and (X^r, x_{l_0+1}). Thus the required subsequence is obtained in case (b).

If, however

(b″) $K_1 = \Delta^{S'}(x_1, \ldots, x_{l_0}) \le \Phi(n_0, l_0)$,

we then obtain in view of (A.4) and (b)

$$\Delta^S(x_1, \ldots, x_{l_0+1}) \le \Phi(n_0 + 1, l_0) + \Phi(n_0, l_0),$$

which by virtue of the properties (A.3) of the function $\Phi(n, l)$ implies that

$$\Delta^S(x_1, \ldots, x_{l_0+1}) \le \Phi(n_0 + 1, l_0 + 1).$$

This however contradicts the condition of the lemma (i.e., (b″) is impossible).

The lemma is proved. □

We shall now prove the theorem. As was pointed out, $m^S(l) \le 2^l$. Let $m^S(l)$ not be identically equal to 2^l, and let n be the first value of l such that $m^S(l) \ne 2^l$. Then for any sample of size l larger than n, the inequality

$$\Delta^S(x_1, \ldots, x_l) \le \Phi(n, l)$$

is valid. Indeed, otherwise, in view of the lemma's assertion, one could find a subsample $x_1, \ldots, x_n$ such that

$$\Delta^S(x_1, \ldots, x_n) = 2^n, \tag{A.5}$$

which is impossible, since by assumption $m^S(n) \ne 2^n$.

Thus the function $m^S(l)$ either is identically equal to 2^l or is majorized by $\Phi(n, l)$. The theorem is proved. □

Remark. The function $\Phi(n, l)$ can be bounded from the above for $n \le 1$ and $l > n$ as follows:

$$\Phi(n, l) < 1.5 \frac{l^{n-1}}{(n-1)!}. \tag{A.6}$$

Since the relation (A.3) is fulfilled for $\Phi(n, l)$, to prove (A.6) it is sufficient to verify that for $n \ge 1$ and $l > n$ the inequality

$$\frac{l^{n-1}}{(n-1)!} + \frac{l^n}{n!} \le \frac{(l+1)^n}{n!} \tag{A.7}$$

is valid and to verify (A.6) on the boundary, i.e., for $n = 1$ and $l = n + 1$.

The inequality (A.7) is clearly equivalent to

$$l^{n-1}(l + n) - (l + 1)^n \le 0,$$

whose validity follows from Newton's binomial expansion.

It thus remains to verify A.6 on the boundary. For $n = 1$ the verification is direct. Next we shall verify the bound for small values of n and l:

$l = n + 1$	2	3	4	5	6
$\Phi(n, l)$	1	4	11	26	57
$1.5\dfrac{l^{n-1}}{(n-1)!}$	1.5	4.5	12	31.25	81

To check (A.6) for $n \ge 6$ we shall utilize Stirling's formula for an upper bound on $l!$:

$$l! \le \sqrt{2\pi l}\, l^l e^{-l+(12l)^{-1}},$$

whence for $l = n + 1$

$$\frac{l^{n-1}}{(n-1)!} = \frac{(l-1)l^{(l-1)}}{l!} \ge \frac{l-1}{\sqrt{2\pi l}\, l}\, e^{-l+(12l)^{-1}},$$

and furthermore for $l \ge 6$

$$\frac{l^{(n-1)}}{(n-1)!} \ge 0.8 \frac{1}{\sqrt{2\pi l}} e^l.$$

On the other hand, $\Phi(n, l) \le 2^l$ always. Therefore it is sufficient to verify that for $l \ge 6$

$$2^l \le 1.2 \frac{1}{\sqrt{2\pi l}} e^l.$$

Actually it is sufficient to verify the inequality for $l = 6$ (which is carried out directly) since as l increases the right-hand side of the inequality grows faster than the left-hand side (for $l > 2$).

Thus we have seen that either the growth function is identically 2^l, or for some n the equality is violated for the first time (i.e., $m^S(n) \neq 2^n$), and then the growth function is bounded by a polynomial function

$$m^S(l) < 1.5 \frac{l^{n-1}}{(n-1)!}.$$

Therefore in order to estimate the behavior of a growth function it is sufficient to find the smallest n such that on no sequence of length l does the system S induce all possible subsequences.

§A3 The Basic Lemma

Let a sample of size $2l$ be chosen:

$$X^{2l} = x_1, \ldots, x_l, x_{l+1}, \ldots, x_{2l},$$

and the frequencies of occurrence of the event $A \in S$ on the first half sample $x_1, \ldots, x_l$ and on the second half sample $x_{l+1}, \ldots, x_{2l}$ be computed. Denote these frequencies by $v'(A)$ and $v''(A)$ respectively, and consider the deviations of these quantities:

$$\rho_A(x_1, \ldots, x_{2l}) = |v'(A) - v''(A)|.$$

We are interested in the maximal deviation of the frequencies over all events of the class S:

$$\rho^S(x_1, \ldots, x_{2l}) = \sup_{A \in S} \rho_A(x_1, \ldots, x_{2l}).$$

Introduce the notation

$$\pi^S(x_1, \ldots, x_{2l}) = \sup_{A \in S} |v'(A) - P(A)|.$$

Furthermore we shall assume that $\pi^S(x_1, \ldots, x_l)$ and $\rho^S(x_1, \ldots, x_{2l})$ are measurable functions.

The Basic Lemma. *The distributions of the quantities $\pi^S(x_1, \ldots, x_l)$ and $\rho^S(x_1, \ldots, x_{2l})$ are related as follows:*

$$P\{\pi^S(x_1, \ldots, x_l) > \varkappa\} \le 2P\left\{\rho^S(x_1, \ldots, x_{2l}) > \frac{\varkappa}{2}\right\},$$

provided that $l > 2/\varkappa$.

Proof. By definition

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X(2l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X^{2l}),$$

where

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

Taking into account that the space $X(2l)$ of samples of size $2l$ is a direct product of $X_1(l)$ and $X_2(l)$ of half samples of size l, we have the equality

$$\int_{X(2l)} \varphi(x_1, \ldots, x_{2l})\, dX^{2l} = \int_{X_1(l)} \left[\int_{X_2(l)} \varphi(x_1, \ldots, x_{2l})\, dX_2^l\right] dX_1^l$$

for any measurable function $\varphi(x_1, \ldots, x_{2l})$, by Fubini's theorem [28].

Therefore

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X_1(l)} dP(X_1^l) \int_{X_2(l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l)$$

(in the inner integral the first half sample is fixed). Denote by Q the event in the space $X_1(l)$

$$\{\pi^S(x_1, \dots, x_l) > \varkappa\},$$

and bounding the domain of integration, we obtain

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} \geq \int_Q dP(X_1^{2l}) \int_{X_2(l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l). \qquad \text{(A.8)}$$

We now bound the inner integral on the right-hand side of the inequality and denote it by I. Here the sample $x_1, \dots, x_l$ is fixed and is such that

$$\pi^S(x_1, \dots, x_l) > \varkappa.$$

Consequently there exists an $A^* \in S$ such that

$$|P(A^*) - \nu(A^*; x_1, \dots, x_l)| > \varkappa.$$

Then

$$I = \int_{X_2(l)} \theta\left[\sup_{A \subset S} \rho_A(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l) \geq \int_{X_2(l)} \theta\left[\rho_{A^*}(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l).$$

Let, for example,

$$\nu'(A^*; x_1, \dots, x_l) < P(A^*) - \varkappa$$

(the case $\nu'(A^*) \geq P(A^*) + \varkappa$ is dealt with completely analogously). Then in order that the conditions

$$|\nu'(A^*; x_1, \dots, x_l) - \nu''(A^*; x_{l+1}, \dots, x_{2l})| > \frac{\varkappa}{2}$$

may be satisfied, it is sufficient that the relation

$$\nu''(A^*) > P(A^*) - \frac{\varkappa}{2}$$

be fulfilled, whence we obtain

$$I \geq \int_{X_2(l)} \theta\left[\nu''(A^*) - P(A^*) + \frac{\varkappa}{2}\right] dP(X_2^l)$$

$$= \sum_{k/l > P(A^*) - \varkappa/2} C_l^k [P(A^*)]^k [1 - P(A^*)]^{l-k}.$$

As is known, the last sum exceeds $\frac{1}{2}$ provided only that $l > 2/\varkappa$. Returning to (A.8), we obtain that for $l > 2/\varkappa$

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} \geq \frac{1}{2}\int_Q dP(X^l) = \tfrac{1}{2}P\{\pi^S(X^l) > \varkappa\},$$

q.e.d. □

§A4 Derivation of Sufficient Conditions

The following theorem is valid.

Theorem A.2. *The probability that for at least one event in the class S the frequency will deviate from the corresponding probability in an experiment of size l by an amount exceeding $\varkappa$ is bounded by*

$$P\{\pi^S(x_1, \ldots, x_l) > \varkappa\} < 6m^S(2l)e^{-\varkappa^2 l/4}. \qquad \text{(A.9)}$$

Corollary. *In order that the frequency of events in class S shall converge (in probability) to the corresponding probabilities uniformly over the class S, it is sufficient that there exist finite n such that for $l > n$*

$$m^S(l) < 1.5\frac{l^{n-1}}{(n-1)!}.$$

Proof. In view of the basic lemma it is sufficient to bound the quantity

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X(2l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X^{2l}).$$

Consider the mapping of the space $X(2l)$ into itself obtained by a permutation T_i of the elements of the sequence X^{2l}. In view of the symmetry of the definition of the measure, the equality

$$\int_{X(2l)} f(X^{2l})\, dP(X^{2l}) = \int_{X(2l)} f(T_i X^{2l})\, dP(X^{2l})$$

holds for any integrable function $f(X)$. Therefore

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X(2l)} \frac{\sum_{i=1}^{(2l)!} \theta\left[\rho^S(T_i X^{2l}) - \frac{\varkappa}{2}\right]}{(2l)!} dP(X^{2l}), \qquad \text{(A.10)}$$

where the sum is taken over all $(2l)!$ permutations.

First we observe that

$$\theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] = \theta\left[\sup_A |v'(A) - v''(A)| - \frac{\varkappa}{2}\right]$$
$$= \sup_A \theta\left[|v'(A) - v''(A)| - \frac{\varkappa}{2}\right]$$

Clearly if two sets A_1 and A_2 induce the same subsample on the sample $x_1, \ldots, x_l, x_{l+1}, \ldots, x_{2l}$, then

$$v'(A_1; T_i X^{2l}) = v'(A_2; T_i X^{2l}),$$
$$v''(A_1; T_i X^{2l}) = v''(A_2; T_i X^{2l}),$$

and hence

$$\rho_{A_1}(T_i X^{2l}) = \rho_{A_2}(T_i X^{2l})$$

for any permutation T_i. In other words, if two events are equivalent with respect to the sample $x_1, \ldots, x_{2l}$, then deviations of frequencies for these events are the same for all permutations T_i. Therefore if from each equivalence class one chooses one set and forms a finite system S', then

$$\sup_{A \in S} \rho_A(T_i X^{2l}) = \sup_{A \in S'} \rho_A(T_i X^{2l}).$$

The number of events in the system S' is finite and is denoted by $\Delta^{S'}(x_1, \ldots, x_{2l})$. Replacing the sup operation by a summation, we obtain

$$\sup_{A \in S} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right] = \sup_{A \in S'} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right]$$
$$\leq \sum_{A \in S'} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right].$$

These relations allow us to bound the integrand in (A.10):

$$\sup_{A \in S'} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right]$$
$$= \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \sup_{A \in S'} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right] \leq \sum_{A \in S'} \frac{\sum_{i=1}^{(2l)!} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2}\right]}{(2l)!}$$

The expression in the square brackets is the ratio of the number of orderings in the sample (of a fixed composition) such that

$$|v'(A) - v''(A)| > \frac{\varkappa}{2},$$

to the total number of permutations. It is easy to see that this expression is equal to

$$\Gamma^* = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

$$k: \left\{ \left| \frac{k}{l} - \frac{m-k}{l} \right| > \frac{\varkappa}{2} \right\},$$

where m equals the number of elements in the sample $x_1, \ldots, x_{2l}$ belonging to A.

In Section A.5 we bound the expression Γ, with the result that

$$\Gamma^* < 3 \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}.$$

Thus

$$\sum_{A \subset S'} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left[\rho_A(T_i X^{2l}) - \frac{\varkappa}{2} \right] < \sum_{A \in S'} 3 \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}$$

$$= 3\Delta^S(x_1, \ldots, x_{2l}) \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}$$

$$\leq 3m^S(2l) \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}.$$

Substituting this bound into the integral (A.10), we obtain

$$P\left\{ \rho^S(X^{2l}) > \frac{\varkappa}{2} \right\} < 3m^S(2l) \exp\left\{ -\frac{\varkappa^2 l}{4} \right\},$$

whence in view of the basic lemma

$$P\{\pi(X^l) > \varkappa\} < 6m^S(2l) \exp\left\{ -\frac{\varkappa^2 l}{4} \right\}.$$

The theorem is proved. □

PROOF OF THE COROLLARY. Let n exist such that for $l > n$

$$m^S(l) < 1.5 \frac{l^{n-1}}{(n-1)!}.$$

Then clearly

$$\lim_{l \to \infty} P\{\pi^S(X^l) > \varkappa\} < 9 \lim_{l \to \infty} \frac{(2l)^{n-1}}{(n-1)!} \exp\left\{ -\frac{\varkappa^2 l}{4} \right\} = 0,$$

i.e., the uniform convergence in probability is valid. □

The sufficient condition obtained does not depend on the properties of the distribution (the only condition is the measurability of functions π^S and ρ^S), but depends on the inner properties of the system S.

Remark. As it was proved in Section A.2 only if the function $m^S(l)$ is not identically 2^l, there exists n such that for $l > n$

$$m^S(l) < 1.5 \frac{l^{n-1}}{(n-1)!}.$$

Therefore the sufficient condition is always fulfilled when

$$m^S(l) \not\equiv 2^l.$$

§A5 A Bound on the Quantity Γ

We bound the expression

$$\Gamma = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

where k runs over the values satisfying the inequalities

$$\left| \frac{k}{l} - \frac{m-k}{l} \right| > \varkappa, \qquad \max(0, m-l) \leq k \leq \min(m, l),$$

or equivalently the inequalities

$$\left| k - \frac{m}{2} \right| > \frac{\varkappa l}{2}, \qquad \max(0, m-l) \leq k \leq \min(m, l),$$

and l and $m \leq 2l$ are arbitrary positive integers.

We decompose Γ into two summands, $\Gamma = \Gamma_1 + \Gamma_2$.

$$\Gamma_1 = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l}, \quad \text{where } k > \frac{\varkappa l}{2} + \frac{m}{2}.$$

$$\Gamma_2 = \sum_k^k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l}, \quad \text{where } k < \frac{\varkappa l}{2} - \frac{m}{2}.$$

Introduce the notation

$$p(k) = \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l} \tag{A.11}$$

$$q(k) = \frac{p(k+1)}{p(k)} = \frac{(m-k)(l-k)}{(k+1)(l+k+1-m)}, \tag{A.12}$$

where

$$\max(0, m-l) \leq k \leq \min(m, l).$$

Furthermore denote

$$s = \min(m, l), \qquad T = \max(0, m - l);$$

$$d(k) = \sum_{i=k}^{s} p(i).$$

Clearly the relation

$$d(k + 1) = \sum_{i=k+1}^{s} p(i) = \sum_{i=k}^{s-1} p(i + 1) = \sum_{i=k}^{s-1} p(i)q(i) \tag{A.13}$$

is valid. Furthermore it follows directly from (A.12) that for $i < j$, $q(i) < q(j)$, i.e., $q(i)$ is monotonically decreasing. Therefore the inequality

$$d(k + 1) = \sum_{i=k}^{s-1} p(i)q(i) < q(k)\sum_{i=k}^{s} p(i)$$

follows from (A.13). Furthermore by definition of $d(k)$ we have

$$d(k + 1) < q(k)\, d(k).$$

Applying this relation successively, we obtain for arbitrary k and j satisfying the condition $T \le j < k \le s$

$$d(k) < d(j) \prod_{i=j}^{k-1} q(i).$$

Furthermore, since $d(j) \le 1$,

$$d(k) < \prod_{i=j}^{k-1} q(i), \tag{A.14}$$

where j is an arbitrary integer smaller than k.

Set

$$t = k - \frac{m - 1}{2}.$$

Then

$$q(t) = \frac{\dfrac{m + 1}{2} - t}{\dfrac{m + 1}{2} + t} \cdot \frac{\left(l - \dfrac{m - 1}{2}\right) - t}{\left(l - \dfrac{m - 1}{2}\right) + t}.$$

Moreover, as long as $T < k < s$, the inequality

$$|t| < \min\left(\frac{m + 1}{2}, l - \frac{m - 1}{2}\right)$$

is clearly valid.

To approximate $q(k)$ we study the function

$$F(t) = \frac{a-t}{a+t} \cdot \frac{b-t}{b+t},$$

assuming that a and b are both positive.

For $|t| < \min(a, b)$

$$\ln F(t) = \ln(a - t) - \ln(a + t) + \ln(b - t) - \ln(b + t).$$

Furthermore we have

$$\ln F(0) = 0, \qquad \frac{d}{dt}(\ln F(t)) = -\left[\frac{2a}{a^2 - t^2} + \frac{2b}{b^2 - t^2}\right].$$

This implies that for $|t| < \min(a, b)$

$$\frac{d}{dt}(\ln F(t)) \le -2\left[\frac{1}{a} + \frac{1}{b}\right].$$

Correspondingly for $|t| < \min(a, b)$ and $t \ge 0$ the inequality

$$\ln F(t) \le -2\left[\frac{1}{a} + \frac{1}{b}\right]t$$

is fulfilled.

Returning to $q(t)$, we obtain for $t \ge 0$

$$\ln q(t) \le -2\left[\frac{2}{m+1} + \frac{2}{2l - m + 1}\right]t = -8\frac{l+1}{(m+1)(2l-m+1)}t.$$

We now bound

$$\ln\left(\prod_{i=j}^{k-1} q(i)\right),$$

assuming that $(m - 1)/2 \le j \le k - 1$:

$$\begin{aligned}\ln\left(\prod_{i=j}^{k-1} q(i)\right) &= \sum_{i=j}^{k-1} \ln q(i)\\ &\le \frac{-8(l+1)}{(m+1)(2l-m+1)} \sum_{i=j}^{k-1}\left(i - \frac{m-1}{2}\right).\end{aligned}$$

Returning to (A.14), we obtain

$$\ln d(k) < \frac{-8(l+1)}{(m+1)(2l-m+1)} \sum_{i=j}^{k-1}\left(i - \frac{m-1}{2}\right);$$

here j is an arbitrary number smaller than k. Therefore for $k > (m - 1)/2$ one can set $j = (m - 1)/2$ for m odd and $j = m/2$ for m even, obtaining a stronger bound. Next, summing the arithmetic progression, we obtain

$$\ln d(k) < \begin{cases} -\dfrac{4(l+1)}{(m+1)(2l-m+1)}\left(k-\dfrac{m}{2}+1\right)^2 & \text{for even } m, \\ -\dfrac{4(l+1)}{(m+1)(2l-m+1)}\left(k-\dfrac{m-1}{2}+1\right)\left(k-\dfrac{m-1}{2}\right) & \\ & \text{for odd } m. \end{cases}$$

Finally Γ_1 is $d(k)$ for the first integer k such that

$$k-\frac{m}{2}>\frac{\varkappa^2 l}{2},$$

whence

$$\ln \Gamma_1 < -\frac{l+1}{(m+1)(2l-m+1)}\varkappa^2 l^2.$$

In the same manner one can bound Γ_2, since the distribution (A.11) is symmetric with respect to the point $k = m/2$. Thus

$$\Gamma < 2\exp\left\{-\frac{(l+1)\varkappa^2 l^2}{(m+1)(2l-m+1)}\right\}. \tag{A.15}$$

The right-hand side of (A.15) attains its maximum at $m = l$, and consequently

$$\Gamma < 2\exp\left\{-\frac{\varkappa^2 l^2}{l+1}\right\} < 3\exp\{-\varkappa^2 l\}.$$

§A6 A Bound on the Probability of Uniform Relative Deviation

In this section we shall prove

Theorem A.3. *For any p $(1 < p \leq 2)$ the bound*

$$P\left\{\sup_{A\in S}\frac{P(A)-\nu(A)}{\sqrt[p]{P(A)}} > \varkappa\right\} < 8m^S(2l)\exp\left\{-\frac{\varkappa^2}{4}l^{2-(2/p)}\right\} \tag{A.16}$$

is valid.

Proof. Consider two events constructed from a random and independent sample of size $2l$: The event Q_1:

$$Q_1 = \left\{\sup_{A\in S}\frac{P(A)-\nu'(A)}{\sqrt[p]{P(A)}} > \varkappa\right\}$$

and the event Q_2:

$$Q_2 = \left\{ \sup_{A \in S} \frac{|v'(A) - v''(A)|}{\sqrt[p]{v(A) + 1/2l}} > \varkappa \right\},$$

where $v'(A)$ is the frequency of the event A computed from the first half-sample of length l; $v''(A)$ is the frequency of the event A computed from the second half-sample; $v''(A)$ is the frequency of the event computed from the sample of length $2l$.

Observe that in the case $l \leq \varkappa^{-p/(p-1)}$ the theorem is trivial. Accordinly we shall prove the theorem as follows: First we show that for $l > \varkappa^{-p/(p-1)}$ the inequality

$$P(Q_1) < 4P(Q_2)$$

is valid, and then we bound the probability of the event Q_2. Thus we shall prove the lemma:

Lemma A.2. *For $l > \varkappa^{-p/(p-1)}$ the inequality*

$$P(Q_1) < 4P(Q_2) \tag{A.17}$$

is valid.

PROOF. Assume that event Q_1 occurred. This means that there exists A^* such that for the first half sample the inequality

$$P(A^*) - v'(A^*) > \varkappa \sqrt[p]{P(A^*)}$$

is fulfilled. Since $v'(A) \geq 0$, this implies that

$$P(A^*) > \varkappa^{p/(p-1)}.$$

Assume that for the second half sample the frequency of occurrence of event A^* exceeds the probability $P(A^*)$:

$$v''(A^*) > P(A^*).$$

Recall now that $l > \varkappa^{-p/(p-1)}$. Under these conditions event Q_2 will definitely occur.

To show this we bound the quantity

$$\mu = \frac{|v'(A^*) - v''(A^*)|}{\sqrt[p]{v(A^*) + 1/2l}} < \frac{v''(A^*) - v'(A^*)}{\sqrt[p]{v(A^*) + 1/2l}} \tag{A.18}$$

under the conditions

$$\begin{aligned} &v'(A^*) < P(A^*) - \varkappa \sqrt[p]{P(A^*)} \\ &v''(A^*) > P(A^*), \\ &P(A^*) > \varkappa^{p/(p-1)}. \end{aligned}$$

For this purpose we find the minimum of the function

$$T = \frac{x - y}{\sqrt[p]{x + y + c}}$$

in the domain $0 < a \leq x \leq 1, 0 < y \leq b, c > 0$. We have for $p > 1$

$$\frac{\partial T}{\partial x} = \frac{1}{p} \frac{(p - 1)x + (p + 1)y + pc}{(x + y + c)^{(p+1)/p}} > 0,$$

$$\frac{\partial T}{\partial y} = -\frac{1}{p} \frac{(p + 1)x + (p + 1)y + pc}{(x + y + c)^{(p+1)/p}} < 0.$$

Consequently T attains its minimum in the admissible domain for $x = a$ and $y = b$. Therefore the quantity μ will be bounded from below if one replaces $v'(A^*)$ by $P(A^*) - \varkappa\sqrt[p]{P(A^*)}$ and $v''(A^*)$ by $P(A^*)$ in (A.18). Thus

$$\mu > \frac{\varkappa\sqrt[p]{2P(A^*)}}{\sqrt[p]{2P(A^*)} - \varkappa\sqrt[p]{P(A^*)} + 1/l}.$$

Furthermore, since $P(A^*) > \varkappa^{p/(p-1)}$, $l > \varkappa^{-p/(p-1)}$, we have

$$\mu > \frac{\varkappa\sqrt[p]{2P(A^*)}}{\sqrt[p]{2P(A^*)} - \varkappa^{p/(p-1)} + \varkappa^{p/(p-1)}} = \varkappa.$$

Thus if Q_1 occurs and the conditions $P(A^*) \leq v''(A^*)$ and $l > \varkappa^{-p/(p-1)}$ are fulfilled, then Q_2 occurs as well.

Observe that the second half sample is chosen independently of the first and, as is known, for $l > 2/P(A^*)$ the frequency of occurrence of the event A^* exceeds $P(A^*)$ with probability $\frac{1}{4}$. Therefore, provided Q_1 is fulfilled, the event

$$v''(A^*) > P(A^*)$$

occurs with probability exceeding $\frac{1}{4}$ as long as $l > \varkappa^{-p/(p-1)}$. Thus for $l > \varkappa^{p/(p-1)}$

$$P(Q_2) > \tfrac{1}{4}P(Q_1).$$

The lemma is proved. □

Lemma A.3. *For any p $(1 < p \leq 2)$ the bound*

$$P(Q_2) < 2m^S(2l) \exp\left\{-\frac{\varkappa^2}{4} l^{2-(2/p)}\right\}$$

is valid.

Proof. Denote by $R_A(X^{2l})$ the quantity

$$R_A(X^{2l}) = \frac{|v'(A) - v''(A)|}{\sqrt[p]{v(A) + 1/2l}}.$$

Then the estimated probability equals

$$P(Q_2) = \int_{X(2l)} \theta\left[\sup_{A\in S} R_A(X^{2l}) - \varkappa\right] dP(X^{2l}).$$

Here the integration is carried out over the space of all possible samples of size $2l$.

Consider now all possible permutations T_i ($i = 1, 2, \ldots, (2l)!$) of the sequence $x_1, \ldots, x_{2l}$. For each such permutation T_i the equality

$$\int_{X(2l)} \theta\left[\sup_{A\in S} R_A(X^{2l}) - \varkappa\right] dP(X^{2l}) = \int_{X(2l)} \theta\left[\sup_{A\in S} R_A(T_iX^{2l}) - \varkappa\right] dP(X^{2l})$$

is valid. Therefore the equality

$$\int_{X(2l)} \theta\left[\sup_{A\in S} R_A(X^{2l}) - \varkappa\right] dP(X^{2l})$$
$$= \int_{X(2l)} \frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \theta\left[\sup_{A\in S} \quad R_A(T_iX^{2l}) - \varkappa\right] dP(X^{2l})$$

is valid.

Consider now the integrand. Since the sample $x_1, \ldots, x_{2l}$ is fixed, instead of the system of events S one can consider a finite system of events S' which contains one representative for each one of the equivalence classes. Thus the equality

$$\frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \theta\left[\sup_{A\in S} R_A(T_iX^{2l}) - \varkappa\right] = \frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \theta\left[\sup_{A\in S'} R_A(T_iX^{2l}) - \varkappa\right]$$

is valid. Furthermore

$$\frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \theta\left[\sup_{A\in S'} R_A(T_iX)^{2l} - \varkappa\right] < \frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \sum_{A\in S'} \theta[R_A(T_iX^{2l}) - \varkappa]$$
$$= \sum_{A\in S'} \left\{\frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \theta[R_A(T_iX^{2l}) - \varkappa\right\}. \tag{A.19}$$

The expression in the braces is the probability of the deviation of frequencies in two half samples for a fixed event A and a given composition of the complete sample. This probability equals

$$\Gamma = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

where m is the number of occurrences of events A in the complete sample, and k is the number of occurrences of the events in the first half sample; k runs over the values

$$\max(0, m - l) \leq k \leq \min(m, l),$$

$$\frac{\left|\dfrac{k}{l} - \dfrac{m-k}{l}\right|}{\sqrt[p]{\dfrac{m+1}{2l}}} > \varkappa.$$

Denote by $\varkappa'$ the quantity

$$\sqrt[p]{\frac{m+1}{2l}}\,\varkappa = \varkappa'.$$

Using this notation the restrictions become

$$\max(0, m - l) \leq k \leq \min(m, l),$$

$$\left|\frac{k}{l} - \frac{m-k}{l}\right| > \varkappa'. \tag{A.20}$$

In Section A.5 the following bound on the quantity Γ under the restrictions (A.20) was obtained:

$$\Gamma < 2\exp\left\{-\frac{(1+1)(\varkappa')^2 l^2}{(m+1)(2l-m+1)}\right\}. \tag{A.21}$$

Expressing (A.19) in terms of $\varkappa$, we obtain

$$\Gamma < 2\exp\left\{-\frac{\varkappa^2(l+1)l^2}{2(2l-m+1)(m+1)}\left(\frac{m+1}{2l}\right)^{2/p}\right\}.$$

The right-hand side of the inequality attains its maximum at $m = 0$. Thus

$$\Gamma < 2\exp\left\{-\frac{\varkappa^2}{4}\,l^{2-(2/p)}\right\}. \tag{A.22}$$

Substituting (A.22) into the right-hand side of (A.19) and integrating, we have

$$P(Q_2) < 2m^S(2l)\exp\left\{-\frac{\varkappa^2}{4}\,l^{2-(2/p)}\right\}. \tag{A.23}$$

The lemma is thus proved. □

The inequalities (A.17) and (A.23) yield the assertion of the theorem. □