V.V.V'yugin

LECTURE NOTES ON MACHINE LEARNING AND PREDICTION

(DRAFT)

V.V.V'yugin "Lecture notes on Machine Learning and Prediction" 2012. - 354 pp.

These Lecture Notes can serve as an initial introduction to the mathematical foundations of the modern theory of machine learning and game-theoretic predictions. The purpose of this guide is to give an overview of the basic mathematical methods and algorithms that are widely discussed in scientific literature in recent years.

The first part sets out the basis of the statistical theory of machine learning and the problem of classification and regression with support vector machines. In the second part, we consider the problems of universal and competitive online prediction (prediction with expert advice). The third part is devoted to game-theoretic interpretation of probability and predictions.

Contents

Introduction

The main goal of science and real life – getting the right predictions about the future behavior of complex systems on the basis of their past behavior.

Many problems, occurring in the real life, cannot be solved using previously known methods or algorithms. This is for the reason that we are not known mechanisms generating baseline data or the information available for us is insufficient to build a model explained data. As they say, we get the data from "the black box".

Under these conditions, we have no choice but learn available to us sequence of input data and to try to construct predictions improving our scheme in the process of prediction. The approach, in which past data or examples are used for the initial formation and improvement schemes of prediction is called *Supervised machine learning*.

Note two types of machine learning methods: batch and online learning. Under *batch learning*, a part of a data set – *training set* is used for training. Once the method of prediction is determined by the training set, it does not change more in the future. Further this method is used for performing predictions on *a testing set*.

In the second type of learning – *online learning*, the process of training never stops. We produce predictions and provide training permanently in the process of data availability.

Machine learning methods of the first type will be considered in Chapters 1 and 2. These chapters are devoted to statistical theory of machine learning and support vector machines.

Methods of the second type will be studied in Chapter 3 in which a theory of *the well calibrated* forecasting is presented, and in Chapters 4 and 6, where *competitive theory* of prediction or *prediction with expert advice* are studied.

In the statistical theory of machine learning the problems of classification and regression are considered. The learning process starts with selecting a classification or a regression function from the predefined large class of such functions.

When the prediction scheme is specified, we have to assess its capability, ie, the quality of its predictions. We first recall, the process of an online statistical model assessment. We suppose that the observed data is generated by some stationary stochastic process. Given past outcomes, we estimate the parameters of this process and update our prediction rule. In this case, *a risk functional* of a given prediction rule is defined as the mean value of some loss function. This expectation is calculated with respect to the "true" underlying probability distribution generating data. Different prediction rules are compared according to values of the risk functional.

In the statistical theory of machine learning, we also refer to some underlying probability distribution generating data. We assume that each training or test example is generated at random from a fixed but unknown to us probability distribution and that the data is independently and identically distributed (i.i.d.). The first step aside from the classical theory is that the distribution generating the data, we may not be known and we can not estimate its parameters. In this case, the bounds of classification (or regression) errors are distribution independent. We refer to such a bound as to *a generalization error*.

A set of methods for assessing the quality of classification and regression schemes is called *generalization theory*.

In this theory, the estimates of classification error are computed, provided that the training was carried out on a random training sample large enough and its resulting classification function agreed with the training set.

The most important parameter of such an assessment is *capacity* or dimension of a class of classification functions. Usually in assessing

of classification errors the length of a training set and the capacity of a class of classification functions are competing – the longer the training set the greater the capacity of a class of hypotheses can be used.

Methods for computing the generalization error and dimension theory of classes of functions are discussed in Chapter 1.

Chapter 2 is devoted to construction of algorithms of classification and regression. Basically, these are algorithms that use Support Vector Machines.

The theory of sequential prediction (Chapter 3) does not use the hypotheses on existing of stochastic mechanism generating the data.

Observed outcomes can be generated by an unknown mechanism, which can be either deterministic or stochastic, or even adversatively adaptive to our predictions (ie, it can use our past predictions for generation of the next outcome).

A natural problem arises – how to evaluate the quality of our predictions. The risk function in the form of the mathematical expectation can not be used, since a probabilistic model is undefined. We shall change them for specific tests that evaluate the disagreement between the predictions and the corresponding outcomes. We consider the calibration tests. The purpose of a forecasting algorithm – to output predictions that passed all tests of calibration.

The basic principles of the prediction with expert advice are discussed in Chapter 4. In the theory of prediction with expert advice, the effectiveness of any forecasting algorithm or *a learner* is evaluated in the form of competition with the a set of expert methods or just the experts. Some loss function measuring the conformity of the prediction and outcome is fixed. The set of experts can be finite or infinite. Experts are presented by the different forecasting methods, stochastic or deterministic. The experts offer their predictions before the outcome be presented. The learner observing these predictions and the cumulative losses of experts, outputs its prediction. After that a new outcome appear and the experts and the learner suffer their losses. The quality of the learner is evaluated in the worst case, namely, in the form of the difference between the cumulative loss of the learner and cumulative losses of the experts. The minimal value

of this difference is called *regret* of the learning algorithm.

In Section 4.1 we study the *weighted majority algorithm* proposed by Littlestone and Warmuth that is the first known algorithm of this type. In Section 4.2 we consider the decision-theoretic online allocation algorithm proposed by Freund and Schapire which solves the problem of prediction with expert advice in most general setting, where only losses of of experts are known. We study in Section 4.4 an exponential weighted forecaster that computes its own predictions using the method of exponential weighting of predictions made by the experts.

In Section 4.7 the method of decision-theoretic online learning is applied for constructing the *boosting* algorithm. In this case "a weak" learning algorithm that performs just slightly better than random guessing can be "boosted" into an arbitrarily accurate "strong" learning algorithm. The famous algorithm AdaBoost of Freund and Shapire is presented.

In Chapter 5 we consider the Vovk's aggregating algorithm which is in some sense equivalent to all previously considered algorithms. The aggregating algorithm be applied to some specific loss functions has significantly smaller regret than algorithms of the exponential weighting. Some applications of this algorithm will be considered, in particular, for multidimensional online regression and for Cover's universal portfolio selection.

Some elements of the classical game theory are studied in Chapter 6. We consider the two-person zero-sum matrix game and prove the von Neumann minimax theorem. This proof is in the theory of machine learning style, it uses the exponential weighting algorithm. In this chapter, we introduce also the notion of the Nash equilibrium and the notion of the correlated equilibrium of Aumann.

Chapter 7 is devoted to a new game-theoretic approach to probability theory proposed by Vovk and Shafer [26]. In this framework, the games of prediction are presented. The forecasting process is considered as a repetitive perfect-information game with players: *Forecaster* generating predictions and *Nature* generating outcomes. The restrictions for the players are regulated by a protocol of the game. Any player receives gain or suffer loss at each round of the game. A

player wins if his cumulative gain increases indefinitely as the number of rounds increases. An additional requirement is that the playing strategy of *Forecaster* should be "defensive". This means, that starting with some initial capital the player never incur debt in the process of the game. An auxiliary player *Skeptic* determines the goal of the game. We show that *Skeptic* can "force" infinite and finite game-theoretic versions of the law of large numbers from the probability theory.

Under this approach, the problem of universal predictions discussed in Chapter 3 is formulated in a natural way. It is shown that *Skeptic* can force *Forecaster* to output forecasts well calibrated on a sequence of outcomes generated by *Nature* independently of that strategy *Nature* uses. These the well calibrated forecasts are constructed using the minimax theorem.

In Chapter 8 the more advanced problems of the game theory are studied. The basis of the proposed theory is the famous Blackwell approachability theorem. The Blackwell's theorem is a generalization of the minimax theorem for vector-valued payoff functions. Using this theorem, we construct the well calibrated forecasts for the case of an arbitrary finite outcome space.

We apply the Blackwell approachability theorem to prove that empirical frequencies of play in any normal-form game with finite strategy sets converges to a set of correlated equilibrium if each player chooses his gamble as the best response to the well calibrated forecasts of the gambles of other players.

We have mapped out various one-semester courses on machine learning and prediction based on this guide. The first of them – "Statistical learning and support vector machines", can be based on Chapters 1 and 2. The second course – "Prediction with experts advice", can be based on Chapters 4 and 5. The third course – "Games on prediction", can be based on the material of Chapters 3, 6, 7 and maybe on Chapter 8. Chapters 6 and 8 can also be a base for an advanced mini-course "Blackwell approachability theorem and its applications".

Part I

Statistical Learning Theory

Chapter 1

Generalization theory

1.1. Classification

1.1.1. Bayes classifier

Given a probability distribution generating pairs (x, y), one may easily construct a classifier with minimal probability of error.

Let a pair of random variables (X, Y) taking values in a set $\mathcal{X} \times \{-1, 1\}$ and distributed with respect to a probability distribution P are given such that the a posteriori probabilities of objects x can exist:

$$P\{Y = 1|X = x\} = cP(x|Y = 1)P\{Y = 1\},$$
$$P\{Y = -1|X = x\} = cP(x|Y = -1)P\{Y = -1\},$$

where

$$c = \frac{1}{P(x|Y = 1)P\{Y = 1\} + P(x|Y = -1)P\{Y = -1\}},$$

Let us denote:

$$\eta(x) = P\{Y = 1|X = x\}.$$

For any classifier $g : \mathcal{X} \to \{-1, 1\}$, the probability of error is defined:

$$\mathrm{err}_P(h) = P\{g(X) \neq Y\}.$$

The Bayes classifier is defined:

$$h(x) = \begin{cases} 1 \text{ if } \eta(x) > \frac{1}{2}, \\ -1 \text{ otherwise.} \end{cases}$$

The following proposition shows that the Bayes classifier has a minimal probability of error. This error is called *Bayes error*.

Proposition 1.1. *For any classifier $g : \mathcal{X} \to \{-1, 1\}$:*

$$P\{h(X) \neq Y\} \leqslant P\{g(X) \neq Y\}. \tag{1.1}$$

Proof. For any classifier g, the conditional probability of error given $X = x$ is defined:

$$\begin{aligned} P\{g(X) \neq Y | X = x\} = \\ = 1 - P\{g(X) = Y | X = x\} = \\ = 1 - (P\{Y = 1, g(X) = 1 | X = x\} + \\ + P\{Y = -1, g(X) = -1 | X = x\}) = \\ = 1 - (1_{g(x)=1} P\{Y = 1 | X = x\} + \\ + 1_{g(x)=-1} P\{Y = 0 | X = x\}) = \\ = 1 - (1_{g(x)=1}\eta(x) + 1_{g(x)=-1}(1 - \eta(x))\}), \end{aligned}$$

where for any predicate $R(x)$ we write $1_{R(x)}(x) = 1$ if $R(x)$ is true and $1_{R(x)}(x) = 0$ otherwise.

The similar equalities hold for the classifier $h(x)$.

Notice that $1_{g(x)=-1} = 1 - 1_{g(x)=1}$ for any classifier g. By definition of Bayes classifier, for all $x \in \mathcal{X}$:

$$\begin{aligned} P\{g(X) \neq Y | X = x\} - P\{h(X) \neq Y | X = x\} = \\ = \eta(x)(1_{h(x)=1} - 1_{g(x)=1}) + \\ + (1 - \eta(x))(1_{h(x)=-1} - 1_{g(x)=-1}) = \\ = (2\eta(x) - 1)(1_{h(x)=1} - 1_{g(x)=1}) \geqslant 0. \end{aligned}$$

Integrating both sides of this inequality by x, we obtain the needed inequality (1.1). $\triangle$

To use the Bayes classifier in practice we have to know the a posteriori distribution $\eta(x)$ that is defined by the distribution $P(x, y)$

generating pairs (x, y). But statistical parameters of mechanisms generating data is usually hard to recover and even they fully unknown to us.

The focus of the theory (and practice) of classification is to construct classifiers $g(x)$ whose probability of error is as close to the Bayes error as possible. Obviously, the whole arsenal of traditional parametric and nonparametric statistics may be used to attack this problem. However, the high-dimensional nature of many of the new applications (such as image recognition, text classification, microbiological applications, etc.) leads to territories beyond the reach of traditional methods. Most new advances of statistical learning theory aim to face these new challenges.

In what follows the introduction of new probability distribution free techniques of handling high-dimensional problems such as boosting and support vector machines have revolutionized the practice of pattern recognition.

1.1.2. Problem setting

Machine Learning theory solves the problems of prediction of a future evolution of complex systems in case where no information is given about mechanisms defining this evolution.

In this books we consider two classes of problems of statistical learning theory: *classification problems* and *regression problems*.

The problem of pattern classification is about guessing or predicting the unknown class of an observation. An observation is often a collection of numerical and/or categorical measurements represented by a n-dimensional vector x that in some cases may be a numerical representation of a curve or of an image.

In general case we simply assume that $x \in \mathcal{X}$, where $\mathcal{X}$ is some abstract measurable object space equipped with an σ-algebra of Borel sets. The unknown nature of the observation is called a class. It is denoted by y and in the simplest case takes values in a finite set D.

In what follows we consider a binary case where $D = \{-1, +1\}$. The reason is simplicity and that the binary problem already captures many of the main features of more general problems. Even though

there is much to say about multiclass classification, this lecture notes does not cover this increasing field of research.

First, we recall the classical setting of PAC-learning model proposed by Valliant [32].

PAC-learning. To formalize the learning problem, we introduce a probabilistic setting. We assume that each training or test example x is generated at random from a fixed but unknown distribution P and that the data is independently and identically distributed (i.i.d.). We next assume that each label is generated by a fixed but unknown target concept $c \in C$. In other words, we assume that the label of x is $c(x)$

Assume that some hypothesis h is defined using a random sample generated by the distribution P. By this definition h can be considered as a random variable. We then define the error of hypothesis h with respect to the target c as

$$\mathrm{err}_P(h) = P\{h(x) \neq c(x)\}.$$

The function $\mathrm{err}_P(h)$ is a random variable, since h is defined by a random sample.

Our goal will be to find a hypothesis h such that the probability that $\mathrm{err}(h)$ is large is small. In other words, we would like to claim that h is probably approximately correct. The "approximately" can be quantified through an accuracy parameter ϵ. In particular, since we will generally not have enough data to learn the target c perfectly, we require only that $\mathrm{err}(h) \leqslant \epsilon$. The "probably" can be quantified through a confidence parameter δ. We can never rule out the unlucky event that we draw an unrepresentative training set and are unable to learn a good approximation of c with the data we have. We instead require that we are able to learn a good approximation with high probability. In particular, we require that $\mathrm{err}(h) \leqslant \epsilon$ with probability at least $1 - \delta$. This leads to the following definition of PAC learning:

An algorithm A PAC-learns a concept class C using a hypothesis class H if for any $c \in C$, for any distribution P over the input space, for any $\epsilon \in (0, 1/2)$ and $\delta \in (0, 1/2)$, given access to a polynomial (in $1/\epsilon$ and $1/\delta$) number of examples drawn i.i.d. from P and labeled by c, the algorithm A outputs a function $h \in H$ such that $\mathrm{err}(h) \leqslant \epsilon$ *with probability at least* $1 - \delta$.

In this lecture notes, we consider a slightly different model considered in the statistical learning theory. We give up the idea of the concept and we assume that the pairs (x, y) of objects x and their labels y are generated i.i.d. by some probability distribution on $\mathcal{X} \times D$.

Strongly speaking, the pairs (x, y) are realizations of a pair of random variables (X, Y) distributed with a density $P(x, y)$.

A classifier, or *a classification rule*, or *a hypothesis*, is a function $h : \mathcal{X} \to D$ that defines a partition of objects from the set $\mathcal{X}$. In binary case the classifier $h : \mathcal{X} \to D$ is also called *an indicator* function. A set $\mathcal{H}$ of hypotheses is often called a hypotheses class.

In some cases, the indicator function h is defined using some real function valued $f : \mathcal{X} \to \mathcal{R}$ and a threshold $r \in \mathcal{R}$:

$$h(x) = \begin{cases} 1 \text{ if } f(x) \geqslant r, \\ -1 \text{ otherwise.} \end{cases}$$

We measure the performance of a classifier h by its probability of error. Let a pair (x, y) be a realization of a random variable (X, Y). The classifier h errs on x if $h(x) \neq y$. We evaluate the performance of a function h by its *classification error* that is the probability of error

$$\mathrm{err}_P(h) = P\{h(x) \neq y\} = P\{(x, y) : h(x) \neq y\}.$$

The function $\mathrm{err}_P(h)$ is also called *a risk-functional*.

Main goal of the classification task is to construct a classifier $h \in \mathcal{H}$ with minimal probability of error $\mathrm{err}_P(h)$.

A simple and natural approach to the classification problem is to consider a class $\mathcal{H}$ of classifiers $h : \mathcal{X} \to \{-1, +1\}$ and use data-based estimates of the probabilities of error $\mathrm{err}_P(h)$ to select a classifier from the class. The most natural choice to estimate the probability of error $\mathrm{err}_P(h)$ is the error count on *a training sample* $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$, where $x_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$ for all $i = 1, \ldots, l$.

We use the assumption that the *ordered sample*

$$S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$$

is generated by some i.i.d. source. This means that some probability distribution P on the probability space $\mathcal{X} \times D$ exists and that all

pairs $(\bar{x}_i, y_i)$, $i = 1, 2, \ldots, l$, are identically and independently distributed according to P. We will also use the corresponding product probability $P^l = P \times P \cdots \times P$ on the product space $(\mathcal{X} \times D)^l$.

In binary case, the training sample S is divided on two subsamples: $S^+ = ((\bar{x}_i, y_i) : y_i = 1)$ – positive examples, and $S^- = ((\bar{x}_i, y_i) : y_i = -1)$ – negative examples.

The empirical error of a classifier h on a training sample S is defined as the portion of mistakes on the sample:

$$\mathrm{err}_S(h) = \frac{1}{l}|\{i : h(x_i) \neq y_i, 1 \leqslant i \leqslant l\}|.$$

Here $|A|$ is the cardinality of a finite set A.

We study these notions in Section 1.2 in a more detail.

In the applications considered below, $\mathcal{X} = \mathcal{R}^n$, where $\mathcal{R}$ is a set of all real numbers, $\mathcal{R}^n$ is the set of all n dimensional vectors, and D is a finite set.

Elements of $\mathcal{R}^n$ are called vectors (points) and denoted by underlined letters: $\bar{x}, \bar{y}, \ \ldots \ \in \mathcal{R}^n$; in coorditates, we write $\bar{x} = (x_1, \ \ldots, \ x_n)'$ – a column-vector. [1]

The standard operations will be considered: summation of vectors

$$\bar{x} + \bar{y} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \cdots \\ x_n + y_n \end{pmatrix}$$

and multiplication of a vector by a real number:

$$\alpha\bar{x} = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \cdots \\ \alpha x_n \end{pmatrix},$$

where $\bar{x} = (x_1, \ \ldots, \ x_n)'$ and $\bar{y} = (y_1, \ \ldots, \ y_n)'$.

[1] Using $'$ sign, we specify the type of the matrix representation of a vector – direct of transposed, but only in cases where this representation is essential.

Another well known operation on $\mathcal{R}^n$ is the dot product: $(\bar{x} \cdot \bar{y}) = x_1y_1 + \cdots + x_ny_n$. The Equlidian norm of a vector $\bar{x}$ of length n is defined:

$$\|\bar{x}\| = \sqrt{(\bar{x} \cdot \bar{x})} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

When solving the multidimensional regression task, we also consider a training set $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$, where y_i is a real number, ie, $D = \mathcal{R}$. The regression task will be considered in Sections 2.8, 2.9, 2.9.1, and in Section 5.9.

1.1.3. Linear classifiers: Perceptron

Perceprton represents some technical model of imagination. [2] This model has two layers: the first receptor layer send a signal to inputs of threshholding elements – neurones of second transforming layer.

Mathematical model of perceptron is described as follows. Let $\mathcal{X}$ be a set of initial descriptions of objects. We call $\mathcal{X}$ the set of *initial features*. A transformation $\bar{y} = \bar{\varphi}(\bar{x})$, that is written in coordinates as $y_i = \varphi_i(\bar{x})$, $i = 1, \ \ldots, \ n$, transforms *initial description* $\bar{x} = (x_1, \ \ldots, \ x_m) \in \mathcal{X}$ of an object to *transformed description* $\bar{y} = (y_1, \ \ldots, \ y_n) \in \mathcal{Y}$ of this object. We suppose that $\mathcal{X} \subseteq \mathcal{R}^m$ and $\mathcal{Y} \subseteq \mathcal{R}^n$ for some m and n. We call $\mathcal{Y}$ the set of *transformed features*.

Perceptron is defined by a homogeneous linear function from the transformed variables:

$$L(\bar{x}) = (\Lambda \cdot \bar{\varphi}(\bar{x})) = \sum_{i=1}^{n} \lambda_i \varphi_i(\bar{x}) = \sum_{i=1}^{n} \lambda_i y_i,$$

where real numbers λ_i are called weights associated with transformed descriptions y_i. Here $(\Lambda \cdot \bar{\varphi}(\bar{x}))$ denotes the dot product of two vectors $\Lambda = (\lambda_1, \ \ldots, \ \lambda_n)$ and $\bar{\varphi}(\bar{x}) = (\varphi_1(\bar{x}), \ \ldots, \ \varphi_n(\bar{x}))$ in the Euclidian space $\mathcal{R}^n$.

We associate with any perceptron *an activation function*:

$$f(\bar{x}) = \sigma\left(\sum_{i=1}^{n} \lambda_i \varphi_i(\bar{x})\right).$$

[2] This section has a rather historical meaning.

Some examples of activation functions:

$$\sigma(t) = \operatorname{sign}(t),$$
$$\sigma(t) = \frac{1}{1+e^{-t}},$$
$$\sigma(t) = \arctan(t),$$

where

$$\operatorname{sign}(t) = \begin{cases} 1 \text{ if } t \geqslant 0, \\ -1 \text{ if } t < 0. \end{cases}$$

In what follows we consider the binary activation function $\sigma(t) = \operatorname{sign}(t)$ defined by a classifier: a vector $\bar{x}$ belongs to the first class if

$$\sum_{i=1}^{n} \lambda_i \varphi_i(\bar{x}) \geqslant 0,$$

a vector $\bar{x}$ belongs to the second class otherwise. [3]

Geometrically, this means that a multi-dimensional manifold

$$\sum_{i=1}^{n} \lambda_i \varphi_i(\bar{x}) = 0 \tag{1.2}$$

is defined in the initial feature space $\mathcal{X}$. This manifold defines a division of the space $\mathcal{X}$ into two subspaces. Objects of the first class locate in one subspace, objects of the second class locate in the second subspace. Such the manifold is called *separating manifold*.

The separating manifold (1.2) corresponds to *a separating hyperplane*:

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

in the transformed feature space $\mathcal{Y}$.

Let an infinite oredered training sample

$$S = ((\bar{y}_1, \epsilon_1), (\bar{y}_2, \epsilon_2), \ \ldots)$$

[3] Somewhat different techniques are needed for perceptrons with sigmoidal activation functions.

in the transformed feature space $\mathcal{Y}$ be given, where $\epsilon_i \in \{-1, +1\}$, $i = 1, 2, \ldots$.

Suppose that a hyperplane *strongly* dividing the sample S on two classes according to values of ϵ_i exists. This means that

$$\epsilon_i(\Lambda \cdot \bar{y}_i) > 0 \tag{1.3}$$

for all i, where $\Lambda = (\lambda_1, \ \ldots, \ \lambda_n)$ is a vector of coefficients of the separating hyperplane.

For technical convenience, we transform the sample S as follows. Define a sequence of vectors $\tilde{y}_1, y_2, \ \ldots$, where

$$\tilde{y}_i = \begin{cases} \bar{y}_i \text{ if } \epsilon_i = 1, \\ -\bar{y}_i \text{ if } \epsilon_i = -1, \end{cases}$$

for all i. Then we can rewrite (1.3) in the form

$$(\Lambda \cdot \tilde{y}_i) > 0$$

for all i. Denote

$$\begin{aligned} \rho(\Lambda) &= \min_i \frac{(\Lambda \cdot \tilde{y}_i)}{|\Lambda|}, \\ \rho_0 &= \sup_{\Lambda \neq \bar{0}} \rho(\Lambda), \end{aligned} \tag{1.4}$$

where $|\Lambda| = \sqrt{\sum_{i=1}^{n} \lambda_i^2}$ is the norm of Λ in the Euclidian space $\mathcal{R}^n$.

Then the condition (1.3) is equivalent to the condition: $\rho_0 > 0$.

Historically, the Rosenblatt's algorithm is the first algorithm for computing a separating hyperplane.

Let an infinite ordered training sample

$$S = (\bar{y}_1, \epsilon_1), (\bar{y}_2, \epsilon_2), \ \ldots$$

be given and let a hyperplane $(\Lambda^* \cdot \bar{y}) = 0$ strongly separating this sample exists. This means that

$$(\Lambda^* \cdot \tilde{y}_i) > 0 \tag{1.5}$$

for all i. Suppose that $|\Lambda^*| = 1$.

We strengthen the condition of the strong separability (1.5): we suppose that a separating threshold $\rho_0 > 0$ exists such that

$$(\Lambda^* \cdot \tilde{y}_i) > \rho_0 \tag{1.6}$$

for all i. We suppose also that the lengths of vectors $\bar{y}_i$ are bounded:

$$\sup_i |\bar{y}_i| = D < \infty.$$

Rosenblatt's algorithm

We will learn the perceptron by updating the weight vector Λ at each step of the algorithm.

Let $\Lambda_t = (\lambda_{1,t}, \ \ldots, \ \lambda_{n,t})$ be the current vector of coefficients at step t, $t = 1, 2, \ \ldots$.

We use the sequence of vectors $\tilde{y}_1, \tilde{y}_2, \ \ldots$ defined above.

Define $\Lambda_0 = (0, \ \ldots, \ 0)$.

FOR $t \ = \ 1, \ 2, \ \ldots$

If $(\Lambda_{t-1} \cdot \tilde{y}_t) \geqslant 0$ then define $\Lambda_t = \Lambda_{t-1}$. In case of right classification we do not update the hyperplane.

If $(\Lambda_{t-1} \cdot \tilde{y}_t) < 0$ (a case of wrong classification) then define $\Lambda_t = \Lambda_{t-1} + \tilde{y}_t$. We call this *improving a mistake*.

ENDFOR

The following theorem first proved by A.A. Novikov says that in case where a hyperplane strongly dividing a sample with a positive threshold exists, Rosenblatt's algorithm outputs a strongly separating hyperplane after a finite number of updates.

Theorem 1.1. *If a hyperplane strongly dividing a sample*

$$(\bar{y}_1, \epsilon_1), (\bar{y}_2, \epsilon_2), \ \ldots$$

with a positive threshold exists then the Rosenblatt's algorithm improves a mistake only at most

$$\left\lfloor \frac{D^2}{\rho_0^2} \right\rfloor$$

times. This means that the inequality $\Lambda_t \neq \Lambda_{t-1}$ holds for at most $\left\lfloor \frac{D^2}{\rho_0^2} \right\rfloor$ distinct t. [4] *After that, the separating hyperplane stabilizes and will divide the rest part of infinite sample without mistakes.*

Proof. If Λ_t changes at step t then

$$\|\Lambda_t\|^2 = \|\Lambda_{t-1}\|^2 + 2(\Lambda_{t-1} \cdot \tilde{y}_t) + \|\tilde{y}_t\|^2.$$

Since $(\Lambda_{t-1} \cdot \tilde{y}_t) \leqslant 0$ (a case of wrong classification) and $\|\tilde{y}_t\| \leqslant D$, we obtain:

$$\|\Lambda_t\|^2 \leqslant \|\Lambda_{t-1}\|^2 + D^2.$$

If k such improving happen before step T then

$$\|\Lambda_t\|^2 \leqslant kD^2. \tag{1.7}$$

By (1.6) a unit vector Λ^* exists such that

$$\epsilon_i(\Lambda^* \cdot \tilde{y}_i) \geqslant \rho_0$$

for all i.

Let us estimate $(\Lambda_t \cdot \Lambda^*)$. By definition $(\Lambda_0 \cdot \Lambda^*) = 0$. If the algorithm improves a mistake on step t then

$$(\Lambda_t \cdot \Lambda^*) = (\Lambda_{t-1} \cdot \Lambda^*) + (\Lambda^* \cdot \tilde{y}_t) \geqslant (\Lambda_{t-1} \cdot \Lambda^*) + \rho_0.$$

Otherwise,

$$(\Lambda_t \cdot \Lambda^*) = (\Lambda_{t-1} \cdot \Lambda^*).$$

Therefore, if the algorithm makes k improvements on steps $\leqslant t$ then

$$(\Lambda_t \cdot \Lambda^*) \geqslant k\rho_0.$$

By Cauchy–Shwarz inequality:

$$(\Lambda_t \cdot \Lambda^*) \leqslant \|\Lambda_t\| \cdot \|\Lambda^*\| = \|\Lambda_t\|.$$

Hence,

$$\|\Lambda_t\| \geqslant k\rho_0. \tag{1.8}$$

[4] $\lfloor r \rfloor$ is the integer part of a real number r.

Combining (1.7) and (1.8), we obtain

$$k \leqslant \frac{D^2}{\rho_0^2}.$$

Hence, the total number of improvements is bounded:

$$k \leqslant \left\lfloor \frac{D^2}{\rho_0^2} \right\rfloor$$

Theorem is proved. $\triangle$

The drawback of this theorem is that though it gives an upper bound for total number k of improvements but we cannot compute this number in advance.

Multilayer neural networks.

We can combine perceptrons in *multilayer neural networks*. Any node ν in a neural network computes a function

$$f^{\nu}(\bar{x}) = \sigma((\bar{w}^{\nu} \cdot \bar{x}) + b^{\nu}),$$

where σ is an activation function.

Consider a network containing l layers. Let $n_1, \ \ldots, \ n_l$ be numbers of nodes in the layers of this network. Suppose that the upper layer has only one node: $n_l = 1$.

With any jth node of the ith layer of the network a function $f_{i,j}(\bar{x}) = \sigma((\bar{w}^{i,j} \cdot \bar{x}) + b^{i,j})$ is associated, where $\bar{w}^{i,j}, \bar{x} \in \mathcal{R}^{n_{i-1}}$, $b^{i,j} \in \mathcal{R}$ and $n_0 > 0$.

The neural network can be represented by a set of vector-valued functions

$$f_i : \mathcal{R}^{n_{i-1}} \to \mathcal{R}^{n_i},$$

$i = 1, \ \ldots, \ l$, where $f_i = (f_{i,1}, \ \ldots, \ f_{i,n_i})$.

The output of the neural network is defined by a real-valued function that is a composition:

$$f_l \circ f_{l-1} \circ \cdots \circ f_2 \circ f_1.$$

Vectors $\bar{w}^{i,j}$ are called weights associated with the jth node in the layer i.

1.2. Vapnik–Chervonienkis generalization theory

1.2.1. Upper bounds for classification error

The generalization theory presents upper bounds for classification error of a classifier defined using random training sample.

Statistical learning theory uses a hypothesis on existing a probabilistic mechanism generating the observed data. In classification or regression problems, this data are pairs (x_i, y_i) of objects and their labels generating sequentially according to some unknown to us probability distribution. We do not try to find parameters of this distribution. We suppose only that pairs (x_i, y_i) are i.i.d. (independently identically distributed) with respect to this distribution. Methods used in the statistical learning theory are uniform with respect to all probability distributions from this very broad class.

A classifier (or regression function) is constructed by a training sample using methods of optimization. A class of classification functions can be very broad – from the class of all separating hyperplanes in n-dimensional Euclidian space to a class of arbitrary n-dimensional manifolds that are mapped using kernel methods to hyperplanes in m-dimensional spaces, where m much bigger than n. No probability distributions are used in algorithms computing values of these classifiers.

In this section, let $\mathcal{X}$ be a set of objects supplied by an σ-algebra of Borel sets and a probability distribution P. Also, let $D = \{-1, +1\}$ be a set of labels of elements of $\mathcal{X}$.

Let $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$ be a training sample, where $x_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$ for $1 \leqslant i \leqslant l$.

In probabilistic analysis, we suppose that the training sample S is a vector random variable consisting of random variables (x_i, y_i), $i = 1, \ \ldots \ l$.

Let a classifier $h : \mathcal{X} \to \{-1, 1\}$ be given. *A classification error* (risk-functional) is defined:

$$\mathrm{err}_P(h) = P\{(x, y) : h(x) \neq y\},$$

that is the probability of a wrong classification.

The classifier h is *agreed* with a sample

$$S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$$

if $h(x_i) = y_i$ for all $1 \leqslant i \leqslant l$.

A simple and natural approach to the classification problem is to consider a class of classifiers h and use data-based estimates of the probabilities of error $\mathrm{err}_P(h)$ to select a classifier from the class. The most natural choice to estimate the probability of error $\mathrm{err}_P(h)$ is the error count:

$$\mathrm{err}_S(h) = \frac{1}{l}|\{i : h(x_i) \neq y_i, 1 \leqslant i \leqslant l\}|$$

that is called *the empirical error* of the classifier h on a sample S. Here $|A|$ is the cardinality of a finite set A.

The classifier h is agreed with a sample S if $\mathrm{err}_S(h) = 0$.

For any classifier h and $\epsilon > 0$ we have:

$$\begin{aligned} P^l\{S : \mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon\} = \\ = \prod_{i=1}^{l} P\{h(x_i) = y_i\} = \\ = \prod_{i=1}^{l}(1 - P\{h(x_i) \neq y_i\}) = \\ = (1 - \mathrm{err}_P(h))^l \leqslant e^{-l\epsilon}, \end{aligned} \tag{1.9}$$

where P^l is the product probability distribution generated by P. We have used in this derivation the i.i.d. property of random pairs (x_i, y_i).

Let H be a class of classification hypotheses. For case of finite class H, by (1.9), we have:

$$P^l\{S : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon)\} \leqslant |H|e^{-l\epsilon}. \tag{1.10}$$

We can interpreted the bound (1.10) as follows. Let a critical level $\delta > 0$ of accepting an error classifier $h \in H$ agreeing with a sample S be given. Then by (1.10) we can assert that with probability

$\geqslant 1-\delta$ any classifier $h_S \in H$ constructed using a random training sample S and agreeing with this sample has the classification error $\mathrm{err}_P(h) \leqslant \epsilon = \frac{1}{l}\ln\frac{|H|}{\delta}$.

In other words, any classifier h having a classification error $\mathrm{err}_P(h) > \epsilon$, with probability $\geqslant 1-|H|e^{-l\epsilon}$, will not agreed with any random sample of length l.

For an infinite class H of classifiers a similar bound can be obtained using Vapnik–Chervonenkis generalization theory.

In this case the cardinality of a finite class is replaced by *a growth function* of an infinite class:

$$B_H(l) = \max_{(x_1,x_2,\ldots,x_l)} |\{(h(x_1), h(x_2), \ \ldots, \ h(x_l)) : h \in H\}|.$$

We will study this function in the next section.

Main result of Vapnik–Chervonenkis theory is an analogue of the inequality (1.10) for infinite class H:

Theorem 1.2. *For $l > 2/\epsilon$, the following upper bound is valid:*

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon)\} \leqslant 2B_H(2l)e^{-\epsilon l/4} \quad (1.11)$$

Proof. Let $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ if $x \notin A$. Similarly, $1_{h(x)\neq y}(x,y)$ is a random variable which equals 1, if $h(x) \neq y$, and equals 0 otherwise. Evidently:

$$E1_{h(x)\neq y} = \mathrm{err}_P(h),$$

where E is the mathematical expectation by the measure P. By definition:

$$\mathrm{err}_S(h) = \frac{1}{l}\sum_{i=1}^{l} 1_{h(x_i)\neq y_i}$$

is the frequency of mistakes on a sample S.

The proof of the theorem is based on the following two lemmas.

Lemma 1.1. *Let a class H of classifiers and two random samples S and S' of length l be given. Then, for any $\epsilon > 0$ such that $l > 2/\epsilon$, the inequality:*

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon)\} \leqslant$$
$$\leqslant\ 2P^{2l}\{SS' : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \frac{1}{2}\epsilon)\} \quad (1.12)$$

holds.

Proof. The inequality (1.12) is equivalent to the inequality:

$$P^l\{S : \sup_{h:\mathrm{err}_S(h)=0} \mathrm{err}_P(h) > \epsilon\} \leqslant$$
$$\leqslant 2P^{2l}\{SS' : \sup_{h:\mathrm{err}_S(h)=0} \mathrm{err}_{S'}(h) > \frac{1}{2}\epsilon\}. \tag{1.13}$$

We will prove the inequality (1.13). For any sample S from the left side of the inequality (1.13), denote by h_S some classifier from H such that $\mathrm{err}_S(h_S) = 0$ and $\mathrm{err}_P(h_S) > \epsilon$. By definition h_S is a random variable.

The following inequality is valid by definition of its terms: [5]

$$1_{\mathrm{err}_S(h_S)\,=\,0\&\mathrm{err}_P(h_S)\,>\,\epsilon} 1_{\mathrm{err}_P(h_S)\,-\,\mathrm{err}_{S'}(h_S)\,\leqslant\,\frac{1}{2}\epsilon} \leqslant$$
$$\leqslant 1_{\mathrm{err}_S(h_S)\,=\,0\&\mathrm{err}_{S'}(h_S)>\frac{1}{2}\epsilon}. \tag{1.14}$$

Integrating both sides of the inequality (1.14) by the sample S', we obtain a new inequality:

$$1_{\mathrm{err}_S(h_S)\,=\,0\&\mathrm{err}_P(h_S)\,>\,\epsilon} P^l\{S' : \mathrm{err}_P(h_S) - \mathrm{err}_{S'}(h_S) \leqslant \frac{1}{2}\epsilon\} \leqslant$$
$$\leqslant P^l\{S' : \mathrm{err}_S(h_S) = 0 \& \mathrm{err}_{S'}(h_S) > \frac{1}{2}\epsilon\} \tag{1.15}$$

depending on the sample S.

Using properties of the binomial distribution, we obtain:

$$P^l\{S' : \mathrm{err}_P(h_S) - \mathrm{err}_{S'}(h_S) \leqslant \frac{1}{2}\epsilon\} =$$
$$= P^l\{S' : \mathrm{err}_{S'}(h_S) \geqslant \mathrm{err}_P(h_S) - \frac{1}{2}\epsilon\} =$$
$$= \sum_{\{k:k/l\geqslant p-\epsilon/2\}} \binom{l}{k} p^k (1-p)^{n-k} > \frac{1}{2} \tag{1.16}$$

[5] Here $1_{\mathrm{err}_S(h_S)=0\&\mathrm{err}_P(h_S)>\epsilon}(S) = 0$ if S does not belong to the left side of the inequality (1.13). Also, $1_{\mathrm{err}_S(h_S)=0\&\mathrm{err}_{S'}(h_S)>\frac{1}{2}\epsilon}(SS') = 0$ if SS' does not belong to the right side of the inequality (1.13).

for $l > 2/\epsilon$, where $p = \mathrm{err}_P(h_S)$.

Indeed, if $l > 2/\epsilon$ then $p - \epsilon/2 < p - 1/l$. Therefore, it is sufficient to prove that

$$\sum_{\{k:k/l \geqslant p-1/l\}} \binom{l}{k} p^k (1-p)^{n-k} = \sum_{\{k:k \geqslant lp-1\}} \binom{l}{k} p^k (1-p)^{n-k} > \frac{1}{2}.$$

This inequality is equivalent to the inequality

$$\sum_{\{k:k<lp-1\}} \binom{l}{k} p^k (1-p)^{n-k} < \frac{1}{2}$$

which is a corollary from the well known fact that the mediane of binomial distribution is equal to the integer number closest to lp.

Combining the inequalities (1.16) and (1.15), we obtain:

$$\begin{aligned} 1_{\mathrm{err}_S(h_S)=0 \& \mathrm{err}_P(h_S)>\epsilon} \leqslant \\ \leqslant 2P^l\{S' : \mathrm{err}_S(h_S) = 0 \& \mathrm{err}_{S'}(h_S) > \frac{1}{2}\epsilon\}. \end{aligned} \tag{1.17}$$

Computing the average by S, we obtain:

$$\begin{aligned} P^l\{S : \mathrm{err}_S(h_S) = 0 \& \mathrm{err}_P(h_S) > \epsilon\} \leqslant \\ \leqslant 2P^{2l}\{SS' : \mathrm{err}_S(h_S) = 0 \& \mathrm{err}_{S'}(h_S) > \frac{1}{2}\epsilon\} \leqslant \\ \leqslant 2P^{2l}\{SS' : \sup_{h:\mathrm{err}_S(h)=0} \mathrm{err}_{S'}(h) > \frac{1}{2}\epsilon\}. \end{aligned} \tag{1.18}$$

From this (1.13) follows. Lemma is proved. $\triangle$

Lemma 1.2. *For any random samples S and S' of length l and $\epsilon > 0$, the probability of that a classifier $h \in H$ is agreed with S and makes more than ϵl mistakes on S' is bounded:*

$$P^{2l}\{SS' : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \epsilon)\} \leqslant B_H(2l)e^{-\epsilon l/2}.$$

Proof. Define a function η that when fed with a sample $SS' = ((x_1, y_1), \ \ldots, \ (x_{2l}, y_{2l}))$ of length $2l$ outputs a bag (or multiset)

whivh is the set is a set of all pairs from the sample SS' with their multiplicities:

$$\eta(SS') = \{((x_1, y_1), k_1), \ \ldots, \ ((x_L, y_L), k_L)\},$$

where k_i is the total number of occurrences of the pair (x_i, y_i) in the sample SS', $i = 1, \ \ldots, \ L$, L is the total number of distinct pairs (x_i, y_i) in the sample SS'. By definition $k_1 + \cdots + k_L = 2l$.

Unlike the sample, the bag is a non ordered object – a set. The probability measure P^{2l} on samples of length $2l$ induces a measure $\hat{P}$ on bags:

$$\hat{P}(\Xi) = P^{2l}\{SS' : \eta(SS') \in \Xi\},$$

where Ξ is any set of bags Υ.

Let us fix some bag Υ for samples of length $2l$. We fix also some classifier h.

For any double sample $SS' = ((x_1, y_1), \ \ldots, \ (x_{2l}, y_{2l}))$ define a binary sequence $\epsilon_1, \ \ldots, \ \epsilon_{2l}$ representing all classification errors of h on SS':

$$\epsilon_i = \begin{cases} 1 \text{ if } h(x_i) \neq y_i, \\ -1 \text{ if } h(x_i) = y_i, \end{cases}$$

where $i = 1, \ \ldots, \ 2l$.

Since the classification mistakes are distributed according to the Bernoulli distribution with probability of error $P\{h(x) \neq y\}$, any two sequences $\epsilon_1, \ \ldots, \ \epsilon_{2l}$ and $\epsilon'_1, \ \ldots, \ \epsilon'_{2l}$ defined by two samples with the same bag Υ are equiprobable. [6]

Therefore, given a bag Υ, the probability of that for some double random sample SS' such that $\eta(SS') = \Upsilon$ the classifier h makes $m \geqslant \epsilon l$ mistakes and all these mistakes are located in the second half

[6] These probabilities are defined by the binomial distribution and equal to $\binom{2l}{k} p^k (1-p)^{2l-k}$, where $p = P\{h(x) \neq y\}$ and k is the number of ones (the number of mistakes) in the sample.

S' of the sample SS' is bounded by the total number of such samples:

$$\frac{\binom{l}{m}}{\binom{2l}{m}} = \frac{l!}{(l-m)!m!} \cdot \frac{(2l-m)!m!}{(2l)!} = \\ = \frac{(2l-m)\dots(l-m+1)}{2l\dots(l+1)} \leqslant \\ \leqslant \left(1-\frac{m}{2l}\right)^l \leqslant \left(1-\frac{\epsilon}{2}\right)^l < e^{-\epsilon l/2}. \tag{1.19}$$

Now, let the classifier h is not fixed and takes any value from the class H. The total number of all projections of functions $h \in H$ on the set $\{x_1, \dots, x_{2l}\}$ does not exceed the cardinality of the set $\{(h(x_1), h(x_2), \dots, h(x_{2l}) : h \in H\}$ consisting from binary sequences of length $2l$.

For any N the total number of all projections of functions $h \in H$ on sets of N objects is upper bounded by *growth function* of the class H:

$$B_H(N) = \max_{(x_1,x_2,\dots,x_N)} |\{(h(x_1), h(x_2), \dots, h(x_N)) : h \in H\}|.$$

Evidently, $B_H(N) \leqslant 2^N$. Tight estimates of the growth functions for different classes H of classifiers will be given in the next section.

By definition of the growth function the total number of all projections of functions $h \in H$ on the set $\{x_1, \dots, x_{2l}\}$ of all objects from the double sample SS' is less or equal to $B_H(2l)$.

Then the conditional probability of that a classifier $h \in H$ makes $\geqslant \epsilon l$ mistakes on a double sample SS' such that $\eta(SS') = \Upsilon$ and all these mistakes are located in the second half S' of this sample is bounded:

$$P^{2l}\{SS' : (\exists\, h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \epsilon) | \eta(SS') = \Upsilon\} \leqslant \\ \leqslant B_H(2l)e^{-\epsilon l/2}.$$

The left part of this inequality is a random variable that is a function from the bag Υ. The right part is independent from the bag Υ.

Integrating this inequality by the measure $\hat{P}$ on bags Υ, we obtain the unconditional inequality:

$$P^{2l}\{SS' : (\exists\, h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \epsilon)\} \leqslant \\ \leqslant B_H(2l)e^{-\epsilon l/2}.$$

Lemma 1.2 is proved. $\triangle$

Theorem 1.2 follows immediately from Lemmas 1.1 and 1.2.

Theorem 1.2 implies that any classifier h, which has a classification mistake $\mathrm{err}_P(h) > \epsilon$, with probability $\geqslant 1 - 2B_H(2l)e^{-\epsilon l/4}$ will not agreed with a random training sample of length $l > 2/\epsilon$. In the process of learning this classifier will be rejected with probability at least $1 - 2B_H(2l)e^{-\epsilon l/4}$ as a wrong classifier.

Denote $\delta = 2B_H(2l)e^{-\epsilon l/4}$. Then for $0 < \delta < 1$ the inequality $l\epsilon > 2$ holds, ie the assumption of Theorem 1.2 is valid. From this, the following corollary can be easily obtained:

Corollary 1.1. *Assume that a class H of classifiers has a finite VC-dimension d.* [7]

Let a critical level $0 < \delta < 1$ of accepting a wrong classification hypothesis $h \in H$ agreeing with a training sample S be given.

Then with P^l-probability $\geqslant 1 - \delta$ a classifier $h_S \in H$ defined by a training sample S and agreeing with it has a classification error:

$$\mathrm{err}_P(h_S) \leqslant \frac{4}{l}\left(d \ln \frac{2el}{d} + \ln \frac{2}{\delta}\right)$$

for $l \geqslant d$.

These results can be generalized for the case of learning with mistakes. The following two Lemmas 1.3 and 1.4 and Theorem 1.3 can be proved.

Lemma 1.3. *Let a class H of classifiers and two random samples S and S' of length l be given. For any $\epsilon > 0$, if $l > 2/\epsilon$ then the*

[7] The definition of VC-dimension is given in the next Section 1.2.2. The bound $B_H(l) \leqslant \left(\frac{el}{d}\right)^d$ for $l \geqslant d$ is obtained in that section.

following inequality holds:

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_P(h) - \mathrm{err}_S(h) > \epsilon)\} \leqslant$$
$$\leqslant 2P^{2l}\{SS' : (\exists\ h \in H)(\mathrm{err}_{S'}(h) - \mathrm{err}_S(h) > \frac{1}{2}\epsilon)\}.$$

The proof of this lemma is similar to the proof of Lemma 1.1.

Lemma 1.4. *The probability of that the empirical errors of a classifier $h \in H$ on two random samples S and S' of length l differs on $\epsilon > 0$ is bounded:*

$$P^{2l}\{SS' : (\exists h \in H)(\mathrm{err}_{S'}(h) - \mathrm{err}_S(h) > \epsilon)\} \leqslant 2B_H(2l)e^{-2\epsilon^2 l}.$$

The proof of this lemma is similar to the proof of Lemma 1.2 with a more complex combinatorical bounds.

In the following theorem we present an upper bound for the probability of that the difference between classification error and empirical error of some classifier $h \in H$ is more than $\epsilon > 0$.

Theorem 1.3. *The following upper bound is valid:*

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_P(h) - \mathrm{err}_S(h) > \epsilon)\} \leqslant 4B_H(2l)e^{-\epsilon^2 l/2} \quad (1.20)$$

for $l > 2/\epsilon$.

The following corollary gives a bound for the classification error of a classifier $h \in H$ in terms of empirical error of this classifier on training sample.

Corollary 1.2. *Assume that a class H of classifiers has a finite VC-dimension d. Then for any $0 < \delta < 1$ and $h \in H$, with probability $\geqslant 1 - \delta$, the following inequality holds:*

$$\mathrm{err}_P(h) \leqslant \mathrm{err}_S(h) + \sqrt{\frac{2}{l}\left(d \ln \frac{2el}{d} + \ln \frac{4}{\delta}\right)},$$

where $l \geqslant d$.

Notice, that the upper bounds given in Theorems 1.2, 1.3, and in Corollaries 1.1 and 1.2 are too rough and mainly have only theoretical meaning, since VC-dimension d of many natural classes is comparable with the length l of a training sample. A more tight dimension free upper bounds will be given in terms of margin-based performance bounds for classification, Rademacher averages and other capacity measures that will be studied in the following sections.

1.2.2. VC-dimension

In this section we study definition and and properties of the Vapnik–Chervonienkis dimension, shortly, VC-dimension, which characterizes a capacity (complexity) of arbitrary infinite class of indicator functions.

Let $\mathcal{X}$ be an object set and H be an arbitrary class of indicator functions defined on $\mathcal{X}$. Let also $h \in H$. A binary sequence $(h(x_1), \ \ldots, \ h(x_l))$ consisting of elements of the set $\{-1, 1\}$ separates the set $\{x_1, \ \ldots, \ x_l\}$ on two subsets: $\{x_i : h(x_i) = 1\}$ – positive examples, and $\{x_i : h(x_i) = -1\}$ – negative examples.

The set $\{x_1, \ \ldots, \ x_l\}$ is *shattered* by the class H if

$$\{(h(x_1), \ \ldots, \ h(x_l)) : h \in H\} = \{-1, 1\}^l.$$

A growth function of the class H is defined as the maximal number of separations of samples of length l on two subsets by means of classifiers from H:

$$B_H(l) = \max_{(x_1, x_2, \ \ldots, \ x_l)} |\{(h(x_1), h(x_2), \ \ldots, \ h(x_l) : h \in H\}|.$$

As follows from the definition, $B_H(l) \leqslant 2^l$ for all l, and if there exists a sample of length l, that is shattered by H, then $B_H(l) = 2^l$.

The following theorem (Sauer's lemma) is the main result of the theory of VC-dimension: [8]

Theorem 1.4. *For any class H of indicator functions, one of two following conditions hold:*

[8] This result was also obtained independently by Vapnik and Chervonenkis (see Vapnik [33], [34]).

1. *$B_H(l) = 2^l$ for all l, ie, for each l an ordered sample of length l shattered by H exists.*

2. *There exists a sample of maximal length d that is shattered by H. In this case $B_H(l) = 2^l$ for $l \leqslant d$ and*

$$B_H(l) \leqslant \sum_{i=0}^{d} \binom{l}{i} \leqslant \left(\frac{el}{d}\right)^d \tag{1.21}$$

for $l > d$.

In other words, the function $G_H(l) = \ln B_H(l)$ is linear for all l or becomes logarithmic: $O(d \ln l)$ for all $l > d$. For example, it cannot be $O(l^d)$ for $0 < d < 1$.

The number d is called VC-dimension (Vapnik–Chervonienkis dimension). If the case (1) is valid then VC-dimension of the class H is infinite.

Proof. Assume that VC-dimension of a class H of indicator functions is equal to d. Then by definition $B_H(l) = 2^l$ for all $l \leqslant d$.

We will prove the inequality (1.21) using the method of mathematical induction by $l + d$.

For $l = d = 1$ this inequality is valid, since both sides of it are equal to 2.

Assume that this inequality is valid for any sum $< l + d$, in particular, for $l - 1$ and d, and for $l - 1$ and $d - 1$.

Let us prove this inequality for the case where the sample size is equal to l and VC-dimension of a class H is equal to d. Denote

$$h(l, d) = \sum_{i=0}^{d} \binom{l}{i}.$$

We have to prove that for any class H with VC-dimension $\leqslant d$ it holds $B_H(l) \leqslant h(l, d)$ for all l.

Using the standard equality for binomial coefficients:

$$\binom{l}{i} = \binom{l-1}{i} + \binom{l-1}{i-1},$$

we obtain the corresponding equality:

$$h(l,d) = h(l-1,d) + h(l-1,d-1).$$

Let H be a class of indicator functions with VC-dimension d and let $X_1 = \{x_1, x_2, \ldots, x_l\}$ be a set of objects of cardinality l, $X_2 = \{x_2, \ldots, x_l\}$ be the same set where the first element is removed.

Let $H_1 = H_{|X_1}$ be a set of all projection of functions from the class H on the set X_1 and $H_2 = H_{|X_2}$ be a set of all projection of functions from the class H on the set X_2.

Let also, H_3 be a class of functions $f \in H_2$ such that a function $f' \in H$ exists which negates a value of f on the removed object: $f'(x_1) = -f(x_1)$.

It holds $|H_1| = |H_2| + |H_3|$, since the class H_2 differs from the class H_1 by the property: for any indicator functions f and f' from the class H_1 taking different values on the object x_1 (in case where such a function exists) only one function from the class H_2 corresponds.

VC-dimension of the class H_2 is less or equal to d, since H_2 is a subclass of the class H_1.

VC-dimension of the class H_3 is less or equal to $d-1$. Indeed, if some set X of cardinality d is shattered by H_3 then the set $X \cup \{x_1\}$, where x_1 is the removed element, is shattered by the class H_1, since for any function $f \in H_3$ two functions $f, f' \in H_1$ exist such that $f(x_1) = -f'(x_1)$. We have the set $X \cup \{x_1\}$ of cardinality $d+1$ which is shattered by the class H_1. This is a contradiction.

By the induction hypothesis:

$$|H_2| \leqslant h(l-1,d) \text{ and } |H_3| \leqslant h(l-1,d-1).$$

Then:

$$|H_1| = |H_2| + |H_3| \leqslant h(l-1,d) + h(l-1,d-1) = h(l,d).$$

Since X is an arbitrary set, we obtain:

$$B_H(l) \leqslant h(l,d) = \sum_{i=0}^{d} \binom{l}{i}.$$

Therefore, the inequality (1.21) is proved.

For $l > d$, the upper bound:

$$B_H(l) \leqslant \sum_{i=0}^{d} \binom{l}{i} \leqslant \left(\frac{el}{d}\right)^d$$

follows from the following chain of inequalities:

$$\sum_{i=0}^{d} \binom{l}{i} \leqslant \left(\frac{l}{d}\right)^d \sum_{i=0}^{d} \binom{l}{i} \left(\frac{d}{l}\right)^i \leqslant$$

$$\leqslant \left(\frac{l}{d}\right)^d \sum_{i=0}^{l} \binom{l}{i} \left(\frac{d}{l}\right)^i =$$

$$= \left(\frac{l}{d}\right)^d \left(1 + \frac{d}{l}\right)^l < \left(\frac{l}{d}\right)^d e^d = \left(\frac{el}{d}\right)^d. \qquad (1.22)$$

Theorem is proved. $\triangle$

In what follows the objects are n-dimensional vectors from the Euclidian space: $\mathcal{X} = \mathcal{R}^n$, where $n \geqslant 1$.

We will compute VC-dimension of the class $\mathcal{L}$ of all linear classifiers on $\mathcal{R}^n$ that are indicator functions $h(\bar{x}) = \text{sign}(L(\bar{x}))$, where $L(\bar{x})$ is a linear function. Recall that $\text{sign}(r) = 1$ if $r \geqslant 0$ and $\text{sign}(r) = -1$ otherwise.

Linear function is any function

$$L(\bar{x}) = (\bar{a} \cdot \bar{x}) + b,$$

where $\bar{x} \in \mathcal{R}^n$ is a variable vector, $\bar{a} \in \mathcal{R}^n$ is a vector of weights, b is a constant.

If $b = 0$ then the linear classifier $\text{sign}(L(\bar{x})) = \text{sign}(\bar{a} \cdot \bar{x})$ is called homogeneous linear classifier.

Evidently, if some set is separated by a linear classifier then it is strongly separated by this classifier.

Theorem 1.5. *1. VC-dimension of the class of all linear classifiers on $\mathcal{R}^n$ is equal to $n + 1$.*

2. VC-dimension of the class of all homogeneous linear classifiers on $\mathcal{R}^n$ is equal to n.

3. *The growth function of the class of all homogeneous linear classifiers on $\mathcal{R}^n$ satisfies the inequality:*

$$G_{\mathcal{L}}(l) = \ln H_{\mathcal{L}}(l) = \ln\left(2\sum_{i=1}^{n-1}\binom{l-1}{i}\right) < \\ < (n-1)(\ln(l-1) - \ln(n-1) + 1) + \ln 2 \qquad (1.23)$$

for $l > n$.

Proof. First, we prove the item 2. A set of n vectors

$$S = \{\bar{e}_1 = (1, 0, \ \ldots, \ 0), \ \ldots, \ \bar{e}_n = (0, 0, \ \ldots, \ 1)\}$$

is shattered by the class of all homogeneous linear classifiers, since for any its subset $\bar{e}_{i_1}, \ \ldots, \ \bar{e}_{i_k}$ a homogeneous linear classifier $h(\bar{x}) = \text{sign}(L(\bar{x}))$, where $L(\bar{x}) = a_1x_1 + \cdots + a_nx_n$, exists which separates this subset from its complement in S. We define coefficients of $L(\bar{x})$ as follows: $a_{i_j} = 1$ for $1 \leqslant j \leqslant k$ and $a_i = -1$ for all other i. Then $L(\bar{e}_{i_j}) = 1$ for $1 \leqslant j \leqslant k$ and $L(\bar{e}_{i_j}) = -1$ for all other j.

Consider an auxiliary matrix:

$$Z = \begin{pmatrix} z_{1,1}, \ \ldots, \ \mathbf{z_{1,j}}, \ \ldots, \ z_{1,2^{n+1}} \\ \ldots \\ z_{i,1}, \ \ldots, \ \mathbf{z_{i,j}}, \ \ldots, \ z_{i,2^{n+1}} \\ \ldots \\ z_{n+1,1}, \ \ldots, \ \mathbf{z_{n+1,j}}, \ \ldots, \ z_{n+1,2^{n+1}} \end{pmatrix},$$

that is defined by the numbers $z_{i,j} = (\bar{a}_j \cdot \bar{u}_i)$, $i = 1, \ldots, n+1$, $j = 1, \ \ldots, \ 2^{n+1}$.

Suppose that some $n+1$ vectors $\bar{u}_1, \ \ldots, \ \bar{u}_n, \bar{u}_{n+1}$ can be (strongly) shattered by the class of all homogeneous linear classifiers. Then 2^{n+1} weight vectors $\bar{a}_1, \ \ldots, \ \bar{a}_{2^{n+1}}$ exist such that the signs of elements of the jth column of the matrix Z correspond to the jth separation of the set S. Hence, there are all 2^{n+1} possible combinations of these signs that are defined by those columns.

Vectors $\bar{u}_1, \ \ldots, \ \bar{u}_n, \bar{u}_{n+1}$ are located in the n-dimensional space and, hence, they are linearly dependent:

$$\lambda_1\bar{u}_1 + \cdots + \lambda_n\bar{u}_n + \lambda_{n+1}\bar{u}_{n+1} = 0 \qquad (1.24)$$

for some real numbers λ_i, $i = 1, \ldots n$, where $\lambda_i \neq 0$ for some i.

Let us consider the dot products of both sides of the equality (1.24) by the vectors $\bar{a}_j$ for $j = 1, \ \ldots, \ 2^{n+1}$. We obtain 2^{n+1} equalities:

$$\lambda_1(\bar{a}_j \cdot \bar{u}_1) + \cdots + \lambda_{n+1}(\bar{a}_j \cdot \bar{u}_{n+1}) = 0.$$

There exists a column with the same signs as the numbers $\lambda_1, \ \ldots, \ \lambda_{n+1}$. Since $\lambda_i \neq 0$,

$$\lambda_1(\bar{a}_i \cdot \bar{u}_1) + \cdots + \lambda_{n+1}(\bar{a}_i \cdot \bar{u}_{n+1}) > 0$$

for some i. This contradiction proves item 2 of the theorem.

Let us prove the item 1. We will prove that the set:

$$\bar{e}_0 = (0, 0, \ \ldots, \ 0)', \bar{e}_1 = (1, 0, \ \ldots, \ 0)', \ \ldots, \ \bar{e}_n = (0, 0, \ \ldots, \ 1)'$$

of $n + 1$ vectors is strongly shattered by the class of all linear classifiers. For any subset $S = \{\bar{e}_{i_1}, \ \ldots, \ \bar{e}_{i_k}\}$ of this set consider a linear classifier

$$h(\bar{x}) = \text{sign}(a_1 x_1 + \cdots + a_n x_n + b), \ \bar{x} = (x_1, \ \ldots, \ x_n),$$

where $a_{i_j} = 1$ for $1 \leqslant j \leqslant k$, and $a_i = -1$ for all other i, $b = \frac{1}{2}$ if $\bar{e}_0 \in S$ and $b = -\frac{1}{2}$ otherwise. It is easy to verify that $L(\bar{e}_{i_j}) > 0$ for all $1 \leqslant j \leqslant k$ and $L(\bar{e}_{i_j}) < 0$ for all other j.

Suppose that a set of n-dimensional vectors

$$\bar{x}_1 = (x_{1,1}, \ \ldots, \ x_{1,n})', \ \ldots, \ \bar{x}_{n+2} = (x_{n+2,1}, \ \ldots, \ x_{n+2,n})'$$

of cardinality $n + 2$ is strongly shattered by the class of all linear classifiers.

We will prove that the set of $n + 1$-dimensional vectors

$$\bar{x}'_1 = (x_{1,1}, \ \ldots, \ x_{1,n}, 1)', \ \ldots, \ \bar{x}'_{n+2} = (x_{n+2,1}, \ \ldots, \ x_{n+2,n}, 1)' \quad (1.25)$$

of cardinality $n+2$ is strongly shattered by the class of homogeneous linear classifiers.

Consider an arbitrary subset $\bar{x}'_{i_1}, \ \ldots, \ \bar{x}'_{i_k}$ of this set and the corresponding subset $\bar{x}_{i_1}, \ \ldots, \ \bar{x}_{i_k}$ of the set (1.25). Suppose that a hyperplane

$$L(\bar{x}) = a_1 x_1 + \cdots + a_n x_n + b$$

separates subset $\bar{x}_{i_1},\ \ldots,\ \bar{x}_{i_k}$ from other vectors of the set (1.25): $L(\bar{x}_{i_j}) > 0$ for $j = 1,\ \ldots,\ k$ and $L(\bar{x}_i) < 0$ for all other i. Define a linear homogeneous classifier in $n+1$-dimensional space:

$$L'(\bar{x}) = a_1 x_1 + \cdots + a_n x_n + b x_{n+1}.$$

Then $L'(\bar{x}'_i) = L(\bar{x}_i)$ for $i = 1,\ \ldots,\ n+2$. Therefore, the linear homogeneous classifier $L'(\bar{x}')$ separates the corresponding subset $\bar{x}'_{i_1},\ \ldots,\ \bar{x}'_{i_k}$ from its complement in the set

$$\bar{x}'_1 = (x_{1,1},\ \ldots,\ x_{1,n}, 1)',\ \ldots,\ \bar{x}'_{n+2} = (x_{n+2,1},\ \ldots,\ x_{n+2,n}, 1)'.$$

Hence, we construct a subset of $n+1$-dimensional space of cardinality $n+2$ shattered by the class of all homogeneous linear classifiers.

This contradiction with the item 2 proves item 1.

Now we turn to the proof of the item 3. Let vectors $\bar{x}_1,\ \ldots,\ \bar{x}_l$ be given. Consider all separations of these set on two classes by means of hyperplanes $L(\bar{u}) = (\bar{u} \cdot \bar{x})$, where $\bar{u}$ is a weight vector defining the hyperplane and $\bar{x}$ is an argument.

For convenience of presentation, we introduce notations: $\mathcal{R}^n(\bar{u}) = \mathcal{R}^n(\bar{x}) = \mathcal{R}^n$. Using notation $\mathcal{R}^n(\bar{u})$ we emphasize that the main variable in this set is u.

Any vector $\bar{u} \in \mathcal{R}^n(u)$ defines a hyperplane $L(\bar{x}) = (\bar{u} \cdot \bar{x})$ in $\mathcal{R}^n(x)$.

We also consider the dual presentation. A vector $\bar{x} \in \mathcal{R}^n(x)$ defines a hyperplane $L(\bar{u}) = (\bar{x} \cdot \bar{u})$ in the space $\mathcal{R}^n(u)$, and l vectors $\bar{x}_1,\ \ldots,\ \bar{x}_l$ from $\mathcal{R}^n(x)$ define l hyperplanes $X_1, \ldots, X_l$ in the space $\mathcal{R}^n(u)$ trespassing the zero point.

Let $\bar{u} \in \mathcal{R}^n(u)$ be a vector defining a hyperplane $L(\bar{u}) = (\bar{u} \cdot \bar{x})$ in $\mathcal{R}^n(x)$ separating the points $\bar{x}_1,\ \ldots,\ \bar{x}_l$ into two subsets. If one continuously rotates this hyperplane in the space $\mathcal{R}^n(x)$ such that the separation of $\bar{x}_1,\ \ldots,\ \bar{x}_l$ remains in fact, the corresponding trajectory of the vector $\bar{u}$ belongs to the same component of the space $\mathcal{R}^n(u)$.

A component is a set of vectors (points) in the space $\mathcal{R}^n(u)$ bounded by the hyperplanes $X_1,\ \ldots,\ X_l$ defining by the weight vectors $\bar{x}_1,\ \ldots,\ \bar{x}_l$. Any such component corresponds to a variant of separation of the vectors $\bar{x}_1,\ \ldots,\ \bar{x}_l$.

The maximal number of different separations of l vectors $\bar{x}_1, \ldots, \bar{x}_l$ by hyperplanes passing through origin in the space $\mathcal{R}^n(x)$ is equal to the number of different components into which l hyperplanes $X_1, \ldots, X_l$ separate the n-dimensional space $\mathcal{R}^n(u)$.

Let $\Phi(n, l)$ be the maximal number of components into which l hyperplanes $X_1, \ldots, X_l$ can divide the n-dimensional space $\mathcal{R}^n(u)$.

We have $\Phi(1, l) = 2$, since the function $L(x) = ux$ can divide any l points on the line into two classes. Also, $\Phi(n, 1) = 2$, since one hyperplane can divide the space $\mathcal{R}^n(u)$ only on two classes.

Given $l-1$ vectors $\bar{x}_1, \ldots, \bar{x}_{l-1}$ in the space $\mathcal{R}^n(x)$ consider the corresponding $l-1$ hyperplanes $X_1, \ldots, X_{l-1}$ in the space $\mathcal{R}^n(u)$. They divide this space into at most $\Phi(n, l-1)$ components.

Adding a new vector $\bar{x}_l$ to these $l-1$ vectors $\bar{x}_1, \ldots, \bar{x}_{l-1}$, we obtain a new hyperplane X_l in the space $\mathcal{R}^n(u)$. The number of components is increased by the quantity equal to the number of components which are split by the hyperplane X_l. Conversely, any such component makes a trace on X_l. The total number of such traces is the total numbers of all parts into which $l-1$ hyperplanes $X_1, \ldots, X_{l-1}$ divide the hyperplane X_l.

Since the dimensionality of X_l is equal to $n-1$, the number of these traces does not exceed $\Phi(n-1, l-1)$.

Then we obtain the following recurrent equation:

$$\Phi(n, l) = \Phi(n, l-1) + \Phi(n-1, l-1), \tag{1.26}$$

where $\Phi(1, l) = 2$, $\Phi(n, 1) = 2$.

Prove as a problem that the recurrence relation (1.26) has the solution:

$$\Phi(n, l) = \begin{cases} 2^l & l \leqslant n \\ 2\sum\limits_{i=1}^{n-1} \binom{l-1}{i} & l > n. \end{cases}$$

To prove the last inequality from (1.23) and the last inequality from (1.21), we can use the bound $\sum\limits_{i=0}^{n} \binom{l}{i} \leqslant \left(\frac{el}{n}\right)^n$, which holds for all $n \leqslant l$. This bound follows from the chain of equalities and inequalities (1.22). The theorem is proved. $\triangle$

Let us obtain an upper bound of VC-dimension of the class of all multilayer neural networks of a given size with the activation function $\sigma(t) = \text{sign}(t)$.

Let $\mathcal{F}$ be a class of the vector-valued indicator functions defined on $\mathcal{R}^n$.

The growth function of the class $\mathcal{F}$ can be written as

$$B_{\mathcal{F}}(m) = \max_{X \subset \mathcal{R}^n, |X|=m} |\mathcal{F}_{|X}|,$$

where $\mathcal{F}_{|X}$ is the class of function which are restrictions of functions from $\mathcal{F}$ on a finite set X.

The needed bound follows from the following proposition.

Proposition 1.2. *Let $\mathcal{F}^1$ and $\mathcal{F}^2$ be two classes of functions and $\mathcal{F} = \mathcal{F}^1 \times \mathcal{F}^2$ be their Cartesian product. Let also, $\mathcal{G} = \mathcal{F}^1 \circ \mathcal{F}^2$ be the class of functions which are compositions of functions from these classes. Then for any n:*

1. $B_{\mathcal{F}}(m) \leqslant B_{\mathcal{F}_1}(m) \cdot B_{\mathcal{F}_2}(m)$;
2. $B_{\mathcal{G}}(m) \leqslant B_{\mathcal{F}_1}(m) \cdot B_{\mathcal{F}_2}(m)$

Proof. To prove (1) notice that for any X such that $|X| = m$:

$$|\mathcal{F}_{|X}| \leqslant |\mathcal{F}^1_{|X}| \cdot |\mathcal{F}^2_{|X}| \leqslant B_{\mathcal{F}^1}(m) \cdot B_{\mathcal{F}^2}(m).$$

We leave the proof of (2) to the reader. $\triangle$

As noted in section 1.1.3 any neural network can be represented as a set of vector-valued functions:

$$f_i : \mathcal{R}^{n_{i-1}} \to \mathcal{R}^{n_i},$$

where n_i are positive integer numbers and $f_i = (f_{i,1}, \ \ldots, \ f_{i,n_i})$ is the a collection of one-dimensional functions of type $\mathcal{R}^{n_{i-1}} \to \mathcal{R}$, $i = 1, \ \ldots, \ l$.

The output of the neural network is a one-dimensional function which is a composition

$$f = f_l \circ f_{l-1} \circ \cdots \circ f_2 \circ f_1.$$

Let $\mathcal{F}$ be a class of all functions f computed by means of the neural networks, $\mathcal{F}^i$ be a class of vector-valued functions $f_i : \mathcal{R}^{n_{i-1}} \to \mathcal{R}^{n_i}$, and $\mathcal{F}^{i,j}$ be the class of jth components of these compositions.

All functions associated with the nodes of the ith layer are linear threshold functions, and thus, VC-dimension of the class $\mathcal{F}^{i,j}$ is equal to $n_{i-1} + 1$ for all j.

By proposition 1.2, and by Sauer lemma, we have:

$$B_{\mathcal{F}}(m) \leqslant \prod_{i=1}^{l} B_{\mathcal{F}^i}(m) \leqslant$$

$$\leqslant \prod_{i=1}^{l} \prod_{j=1}^{n_i} B_{\mathcal{F}^{i,j}}(m) \leqslant$$

$$\leqslant \prod_{i=1}^{l} \prod_{j=1}^{n_i} \left(\frac{le}{n_{i-1}+1} \right)^{d_{i-1}+1} =$$

$$= \prod_{i=1}^{l} \left(\frac{me}{n_{i-1}+1} \right)^{n_i(n_{i-1}+1)} \leqslant (me)^N,$$

where

$$N = \sum_{i=1}^{l} d_i(d_{i-1}+1)$$

is the total number of all parameters of the neural network.

We now estimate the VC-dimension of the class $\mathcal{F}$. Let m be the maximum of the cardinality of sets shattered by functions from the class $\mathcal{F}$. Then $2^m \leqslant (me)^N$. In order to satisfy this inequality m should be $m = O(N \log N)$. Therefore, VC-dimension of the class $\mathcal{F}$ is bounded by $O(N \log N)$.

1.3. Margin-based performance bounds for classification

We have shown in previous sections that VC-dimension of the class of all linear classifiers is equal to $n + 1$, where n is dimension of the Euclidian space $\mathcal{R}^n$. In practice, the length of a sample can be less

than n, and bounds of the classification error like (1.11) and (1.20) are useless in this case.

By this reasons, Theorem 1.2 and Corollary 1.1 can have only a theoretical meaning. These drawbacks are connected with too poor method used for separation of the data. Separating data with arbitrary small thresholds we loss the predictive performance of our classification algorithms. Also, we do not restrict the space where our training sample is located.

In what follows we will consider methods of separation with a given positive threshold γ and will suppose that the points generating by the probability distributions are located in some ball of a given radius R. Using γ and R as the new parameters, we will define a new dimension free notion of capacity of a functional class.

We obtain new upper bounds of classification error which can have some practical meaning.

1.3.1. Fat-shattering dimension and its applications

Let $\mathcal{F}$ be a class of real valued functions with domain $\mathcal{X}$, $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$ be a sample of length l, and $\epsilon > 0$.

Any function $f \in \mathcal{F}$ defines a classifier:

$$h_f(x) = \begin{cases} 1 \text{ if } f(x) \geqslant 0, \\ -1 \text{ otherwise.} \end{cases}$$

For a function $f \in \mathcal{F}$ we define its margin on an example (x_i, y_i) to be $\gamma_i = y_i f(x_i)$.

The functional margin of a training set $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$ is defined to be:

$$m_S(f) = \min_{i=1,\ldots,\ l} \gamma_i$$

If $\gamma_i > 0$ then the classification by means of f is right. It holds $m_S(f) > 0$ if and only if the function f classifies all examples from the sample S right and with a positive threshold.

A finite set $\mathcal{B}$ of functions is called *ϵ-cover* of a functional class $\mathcal{F}$ on a set $X = \{x_1, \ \ldots, \ x_l\}$ if for any $f \in \mathcal{F}$ a function $g \in \mathcal{B}$ exists such that $|f(x_i) - g(x_i)| < \epsilon$ for all $i = 1, \ \ldots, \ l$.

Define the covering number of a class $\mathcal{F}$ on a set X:

$$\mathcal{N}(\epsilon, \mathcal{F}, X) = \min\{|\mathcal{B}| : \mathcal{B} \text{ is } \epsilon\text{-cover of } \mathcal{F}\}.$$

Define *the covering number* $\mathcal{N}(\epsilon, \mathcal{F}, l)$ of a class $\mathcal{F}$ as the maximum number of all covering numbers of the class $\mathcal{F}$ on the sets X such that $|X| = l$:

$$\mathcal{N}(\epsilon, \mathcal{F}, l) = \max_{|X|=l} \mathcal{N}(\epsilon, \mathcal{F}, X).$$

Denote by $\mathrm{err}_S(f)$ the empirical error of a classifier h_f on a sample $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$. This number is equal to the portion in S of all examples (x_i, y_i) such that $h_f(x_i) \neq y_i$.

Let P be a probability distribution on $\mathcal{R} \times \{-1, 1\}$ generating elements of the sample S. Then the classification mistake of the classifier h_f can be written as:

$$\mathrm{err}_P(f) = P\{h_f(x) \neq y\}.$$

The following theorem is an analogue of Theorem 1.2.

Theorem 1.6. *For any $\epsilon > 0$, $\gamma > 0$, and $l > 2/\epsilon$:*

$$P^l\{S : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_P(f) > \epsilon)\} \leqslant \\ \leqslant 2\mathcal{N}(\gamma/2, \mathcal{F}, 2l)e^{-\epsilon l/4}.$$

The proof of Theorem 1.6 is similar to the proof of Theorem 1.2. We have only to add to the equality $\mathrm{err}_S(f) = 0$ in the right side of the condition (1.12) of Lemma 1.3 the inequality $m_S(f) \geqslant \gamma$. So, we replace Lemma 1.3 on the following lemma:

Lemma 1.5. *For $l > 2/\epsilon$:*

$$P^l\{S : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_P(f) > \epsilon)\} \leqslant \\ \leqslant 2P^{2l}\{S\hat{S} : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_{\hat{S}}(f) > \frac{\epsilon}{2})\}.$$

The proof of this lemma is almost identical to the proof of Lemma 1.1.

The second lemma is an analogue of Lemma 1.2.

Lemma 1.6. *For $l > 2/\epsilon$:*

$$P^{2l}\{S\hat{S} : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_S(f) > \frac{\epsilon}{2})\} \leqslant$$
$$\leqslant \mathcal{N}(\gamma/2, \mathcal{F}, 2l)e^{-\epsilon l/4}.$$

Proof. Consider an $\gamma/2$-cover $\mathcal{B}$ of the class $\mathcal{F}$ of objects of a double sample $S\hat{S}$. Let $g \in \mathcal{B}$ approximates the function $f \in \mathcal{F}$ up to $\gamma/2$. If $m_S(f) \geqslant \gamma$ then $m_S(g) > \gamma/2$. Also, if $\mathrm{err}_S(f) = 0$ and $m_S(f) \geqslant \gamma$ then $\mathrm{err}_S(g) = 0$.

If the function f makes a mistake on an object x_i, ie, $y_i f(x_i) \leqslant 0$, then $y_i g(x_i) < \gamma/2$. Let $\mathrm{err}_{\hat{S}}(\gamma/2, g)$ denotes a portion of all i such that $y_i g(x_i) < \gamma/2$, where x_i locates at the second half of the double sample $S\hat{S}$. This implies the inequality:

$$P^{2l}\{S\hat{S} : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_S(f) > \frac{\epsilon}{2})\} \leqslant$$
$$\leqslant P^{2l}\{S\hat{S} : (\exists\ g \in \mathcal{B})(\mathrm{err}_S(g) = 0 \& m_S(g) \geqslant \frac{\gamma}{2} \& \mathrm{err}_{\hat{S}}(\gamma/2, g) > \frac{\epsilon}{2})\}.$$

The further proof repeats the combinatorial part of the proof of Lemma 1.4. In this part, we bound a portion of variants such that some function $g \in \mathcal{B}$:

- separates the first half S of the double sample SS' without mistakes: $\mathrm{err}_S(g) = 0$, moreover, this is a strong separation with a threshold: $m_S(g) > \gamma/2$;
- at the same time, the function g makes a portion of $\mathrm{err}_{\hat{S}}(\gamma/2, g) > \frac{\epsilon}{2}$ mistakes or has margin bound $\leqslant \gamma/2$ on the second half S' of the double sample.

The rest part of proof coincides with the corresponding part of the proof of Lemma 1.2.

As a result we obtain a bound:

$$P^{2l}\{S\hat{S} : (\exists\ g \in \mathcal{B})(\mathrm{err}_S(g) = 0 \& m_S(g) \geqslant \frac{\gamma}{2} \& \mathrm{err}_{\hat{S}}(\gamma/2, g) > \frac{\epsilon}{2})\} \leqslant$$
$$\leqslant \mathcal{N}(\gamma/2, \mathcal{F}, 2l)e^{-\epsilon l/4}.$$

Lemmas 1.5 and 1.6 imply Theorem 1.6. $\triangle$

Theorem 1.6 implies the following corallary:

Corollary 1.3. *Let a class $\mathcal{F}$ of real functions and numbers $\gamma > 0$, $\delta > 0$ be given. Then for any probability distribution P on $\mathcal{R}^n \times \{-1, 1\}$, with probability $1-\delta$, any function $f \in \mathcal{F}$ with margin bound $m_S(f) > \gamma$ on a random sample S has classification error*:

$$\mathrm{err}_P(f) \leqslant \frac{4}{l}\left(\ln \mathcal{N}(\gamma/2, \mathcal{F}, 2l) + \ln \frac{2}{\delta}\right)$$

for all l.

We define a fat-shattering dimension of a class $\mathcal{F}$ of functions. Let $\gamma > 0$. A set $X = \{x_1, \ \ldots, \ x_l\}$ of objects is called *γ-shattered* if the numbers $r_1, \ \ldots, \ r_l$ exist such that for any subset $E \subseteq X$ a function $f_E \in \mathcal{F}$ exists such that $f_E(x_i) \geqslant r_i + \gamma$ if $x_i \in E$ and $f_E(x_i) < r_i - \gamma$ if $x_i \notin E$ for all i.

A set X is *γ-shattered on a given level r* if $r_i = r$ for all i.

A *fat-shattering* dimension $\mathrm{fat}_\gamma(\mathcal{F})$ of a class $\mathcal{F}$ is equal to cardinality of the maximal γ-shattered set X. The fat-shattering dimension of the class $\mathcal{F}$ depends on the parameter $\gamma > 0$. A class $\mathcal{F}$ has infinite fat-shattering dimension if there are γ-shattered sets of arbitrary big size.

The following theorem is a direct corollary of Theorem 1.10, which will be proved in the Section 1.3.2 below.

Theorem 1.7. *Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to [a, b]$, where $a < b$. For $0 < \gamma < 1$ denote $d = \mathrm{fat}_{\gamma/4}(\mathcal{F})$. Then*

$$\ln \mathcal{N}(\gamma, \mathcal{F}, l) \leqslant 1 + d \ln \frac{2el(b-a)}{d\gamma} \ln \frac{4l(b-a)^2}{\gamma^2}.$$

Theorem 1.7 and Corollary 1.3 imply the following corollary.

Corollary 1.4. *Let $\mathcal{F}$ be a class of real functions with the range $[-1, 1]$, $\gamma > 0$, $\delta > 0$, and P be a probability distribution generating a sample S. Then, with probability $1 - \delta$, any hypothesis $f \in \mathcal{F}$, with the margin bound $m_S(f) \geqslant \gamma$ has a classification error*:

$$\mathrm{err}_P(f) \leqslant \frac{4}{l}\left(d \ln \frac{16el}{d\gamma} \ln \frac{128l}{\gamma^2} + \ln \frac{2}{\delta}\right),$$

where $d = \mathrm{fat}_{\gamma/8}(\mathcal{F})$.

A dimension free upper bound of the fat-dimension can be obtained for the class of all (homogeneous) linear functions with restricted domain in $\mathcal{R}^n$:

Theorem 1.8. *Let $X = \{\bar{x} : |\bar{x}| \leqslant R\}$ be a ball of radius R in the n-dimensional Euclidian space and $\mathcal{F}$ be the class of all homogeneous linear functions $f(\bar{x}) = (\bar{w} \cdot \bar{x})$, where $\|\bar{w}\| \leqslant 1$ and $\bar{x} \in X$. Then*

$$\mathrm{fat}_\gamma(\mathcal{F}) \leqslant \left(\frac{R}{\gamma}\right)^2.$$

Proof. Assume that a set $Y = \{\bar{x}_1, \ \ldots, \ \bar{x}_l\}$ is γ-shattered by the class $\mathcal{F}$ of all homogeneous linear functions with witnesses $r_1, \ldots, r_l$.

Let $\hat{Y}$ be an arbitrary subset of Y. Assume that

$$\sum_{\bar{x}_i \in \hat{Y}} r_i \geqslant \sum_{\bar{x}_i \in Y \setminus \hat{Y}} r_i.$$

By definition a weight vector $\bar{w}$, $\|\bar{w}\| \leqslant 1$ exists such that $(\bar{w} \cdot \bar{x}_i) > r_i + \gamma$ for $\bar{x}_i \in \hat{Y}$ and $(\bar{w} \cdot \bar{x}_i) \leqslant r_i - \gamma$ for $\bar{x}_i \notin \hat{Y}$. Then we have

$$\sum_{\bar{x}_i \in \hat{Y}} (\bar{w} \cdot \bar{x}_i) \geqslant \sum_{\bar{x}_i \in \hat{Y}} r_i + |\hat{Y}|\gamma$$

and

$$\sum_{\bar{x}_i \in Y \setminus \hat{Y}} (\bar{w} \cdot \bar{x}_i) \leqslant \sum_{\bar{x}_i \in Y \setminus \hat{Y}} r_i - |Y \setminus \hat{Y}|\gamma.$$

The difference of these sums is estimated:

$$\sum_{x_i \in \hat{Y}} (\bar{w} \cdot \bar{x}_i) - \sum_{x_i \in Y \setminus \hat{Y}} (\bar{w} \cdot \bar{x}_i) \geqslant \gamma l. \tag{1.27}$$

Using Cauchy–Shwarz inequality for the Euclidian norm, we obtain

$$\sum_{x_i \in \hat{Y}} (\bar{w} \cdot \bar{x}_i) - \sum_{x_i \notin \hat{Y}} (\bar{w} \cdot \bar{x}_i) =$$

$$= \left(\bar{w} \cdot \left(\sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \notin \hat{Y}} \bar{x}_i\right)\right) \leqslant$$

$$\leqslant \left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \notin \hat{Y}} \bar{x}_i \right\| \cdot \|\bar{w}\|. \tag{1.28}$$

By (1.27), (1.28) and $\|\bar{w}\| \leqslant 1$, we obtain a lower bound

$$\left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \notin \hat{Y}} \bar{x}_i \right\| \geqslant \gamma l. \tag{1.29}$$

Assume now that

$$\sum_{\bar{x}_i \in \hat{Y}} r_i < \sum_{\bar{x}_i \in Y \setminus \hat{Y}} r_i.$$

Interchange the sets $\hat{Y}$ and $Y \setminus \hat{Y}$ and obtain

$$\left\| \sum_{x_i \in Y \setminus \hat{Y}} \bar{x}_i - \sum_{x_i \in \hat{Y}} \bar{x}_i \right\| \geqslant \gamma l.$$

Therefore, the inequality (1.29) is valid in both cases.

Continue the proof of of the theorem.

Let $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ be a uniformly distributed random vector of length l such that $\xi_i \in \{-1, 1\}$ for $i = 1 \ , \ldots, \ l$.

The binary vector $\bar{\xi}$ shatters naturally the vectors of Y on two subsets $\hat{Y}$ and $Y \setminus \hat{Y}$.

Let us compute the mathematical expectation of the square norm of the difference (1.29) with respect to the probability distribution generating vector $\bar{\xi}$:

$$E \left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \notin \hat{Y}} \bar{x}_i \right\|^2 = E \left\| \sum_{i=1}^{l} \xi_i \bar{x}_i \right\|^2 =$$

$$= E \sum_{i=1}^{l} \xi_i^2 \|\bar{x}_i\|^2 + 2E \sum_{i,j=1, i \neq j}^{l} \xi_i \xi_j (\bar{x}_i \cdot \bar{x}_j) =$$

$$= E \sum_{i=1}^{l} \|\bar{x}_i\|^2 \leqslant R^2 l.$$

A subset $\hat{Y}$ exists such that the square norm of the difference (1.29) is less or equal than its mathematical expectation:

$$\left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \notin \hat{Y}} \bar{x}_i \right\| \leqslant R\sqrt{l}.$$

Together with the inequality (1.29), this implies $R\sqrt{l} \geqslant \gamma l$. From this $l \leqslant (R/\gamma)^2$ follows. This means that $\mathrm{fat}_\gamma(\mathcal{F}) \leqslant (R/\gamma)^2$. $\triangle$

Substituting the bound of Theorem 1.8 in the bound of Corollary 1.4, we obtain the final theorem:

Theorem 1.9. *Let the classification problem by use of linear homogeneous functions $f(\bar{x}) = (\bar{w} \cdot \bar{x})$, where $\bar{x} \in \mathcal{R}^n$ and $\|\bar{w}\| \leqslant 1$, be considered.*

Let a number $\gamma > 0$ and a probability distribution P concentrated in the ball of the radius R and centered in the origin be given. Let also, a sample $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$ be generated by the probability distribution P. Then, with probability $1 - \delta$, any classification hypothesis f with the margin bound $m_S(f) \geqslant \gamma$ has the classification error:

$$\mathrm{err}_P(f) \leqslant \frac{4}{l}\left(\frac{64R^2}{\gamma^2}\ln\frac{el\gamma}{4R}\ln\frac{128Rl}{\gamma^2} + \ln\frac{2}{\delta}\right). \tag{1.30}$$

For this evaluation, we used the fact that in the inequality (**??**) of Theorem 1.8 instead of d we can take any upper bound of the number $d' = Sdim(\mathcal{F}^{\alpha/2})$. In this case it is convenient to take $d = \frac{64R^2}{\gamma^2}$.

The bounds of Theorems 1.8 and 1.9 serve as a basis of the theory of dimension-free bounds of classification errors for Suppotr Vector Machines presented in Theorem 2.4 of Section 2.6.1.

1.3.2. Covering and Packing numbers

In this section we consider the material of the previous section from a more general position.

Let $(\mathcal{X}, d)$ be a metric space with a metrics $d(x, y)$ which defines the distance between any two elements $x, y \in \mathcal{X}$.

Let $A \subseteq \mathcal{X}$, $B \subseteq A$, and α be a positive number. The set B is called α-*cover* of the set A if for any $a \in A$ an $b \in B$ exists such that $d(a, b) < \alpha$. *A covering number* of the set A is a function:

$$\mathcal{N}_d(\alpha, A) = \min\{|B| : B \text{ is } \alpha\text{-covering of } A\}. \tag{1.31}$$

We say that the set $B \subseteq \mathcal{X}$ is $\alpha-$*separated* if $d(a, b) > \alpha$ for any $a, b \in B$ such that $a \neq b$.

A packing number of the set A is a function:

$$\mathcal{M}_d(\alpha, A) = \max\{|B| : B \subseteq A \text{ is } \alpha\text{-separated}\}. \tag{1.32}$$

The covering number and the packing number are closely related:

Lemma 1.7. *For any $A \subseteq \mathcal{X}$ and $\alpha > 0$*

$$\mathcal{M}_d(2\alpha, A) \leqslant \mathcal{N}_d(\alpha, A) \leqslant \mathcal{M}_d(\alpha, A).$$

Proof. Let M be 2α-separated subset of A and N be α-covering of A. By definition of the set N for any $a \in M$ an $b \in N$ exists such that $d(a, b) < \alpha$. If $a, a' \in M$ are different and $b, b' \in N$ such that $d(a, b) < \alpha$ and $d(a', b') < \alpha$ then b and b' are also different, since if $b = b'$ then $d(a, a') \leqslant d(a, b) + d(b, a') < 2\alpha$. This contradicts to the fact that any two distinct elements of M are within a larger than 2α. From this the inequality $|M| \leqslant |N|$ follows. The left-hand side inequality is proven.

Let M be a maximal under inclusion α-separated subset of A. We shall prove that M is an α-covering of the set A. Suppose it is not. Then an $x \in A$ exists such that there are no elements of M in the ball of the radius α centered in x. Adding x to M, we obtain the strictly larger subset $M \cup \{x\}$ of the set A which is also α-separated. This contradicts to the choice of M. This contradiction proves the right-hand inequality. $\triangle$

The main purpose of this section is to prove Theorem 1.10. To carry it out, we need some development of the dimension theory for functions with a finite number of values.

Let $\mathcal{X}$ be a set and $B = \{0, 1, \ldots, b\}$ be a finite set. Let also, $\mathcal{F} \subseteq B^{\mathcal{X}}$ be a class of functions with domain $\mathcal{X}$ and range in the finite set B. Consider a metrics on $\mathcal{F}$:

$$l(f, g) = \sup_{x \in \mathcal{X}} |f(x) - g(x)|.$$

Any two functions $f, g \in \mathcal{F}$ are said to be *separated* (2-separated) if $l(f, g) > 2$. In other words, an $x \in \mathcal{X}$ exists such that $|f(x) - g(x)| > 2$. A class $\mathcal{F}$ is said to be *pairwise separated* if any two different functions $f, g \in \mathcal{F}$ are separated.

Let $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ be a linear ordered set – a sample – and $\mathcal{F} \subseteq B^{\mathcal{X}}$. We say that the class $\mathcal{F}$ *strongly shatters* the set X if there exists a collection $s = (s_1, \ldots, s_n\}$ of elements of B such that for all $E \subseteq X$ a function $f_E \in \mathcal{F}$ exists such that

$$x_i \in E \Longrightarrow f_E(x_i) \geqslant s_i + 1$$
$$x_i \notin E \Longrightarrow f_E(x_i) \leqslant s_i - 1$$

for all i.

In this case we also say that $\mathcal{F}$ strongly shatters the set X according to s. *The strong dimension* of $\mathcal{F}$, denoted $Sdim(\mathcal{F})$, is the size of a largest strongly shattered set.

We will shift our attention from real valued functions $f : \mathcal{X} \to [0, 1]$ to ones taking values in a finite set by a simple discretization. For any real $\alpha > 0$ define

$$f^{\alpha}(x) = \left[\frac{f(x)}{\alpha}\right]$$

for all x, where $[r]$ is the closest to r integer number such that $|r - [r]| \leqslant \frac{1}{2}$. If the number r is located in the middle of the interval between two integer numbers we define $[r]$ using some tie breaking rule. Define $\mathcal{F}^{\alpha} = \{f^{\alpha} : f \in \mathcal{F}\}$.

Clearly, the range of any function f^{α} is a subset of the set $\{0, 1, \ldots, \lfloor 1/\alpha \rfloor\}$.

The covering number $\mathcal{N}_d(\alpha, A)$ and the packing number $\mathcal{M}_d(\alpha, A)$ were defined by (1.31) and (1.32).

Let us define a specific metrics on the class $\mathcal{F}$ connected with the set $X = \{x_1, \ldots, x_n\}$:

$$l_X(f, g) = \max_{1 \leqslant i \leqslant n} |f(x_i) - g(x_i)|.$$

Consider the corresponding covering and packing numbers:

$$\mathcal{N}(\alpha, \mathcal{F}, X) = \mathcal{N}_{l_X}(\alpha, \mathcal{F}),$$
$$\mathcal{M}(\alpha, \mathcal{F}, X) = \mathcal{M}_{l_X}(\alpha, \mathcal{F}).$$

The following lemma relates the combinatorial dimensions and packing numbers of the classes $\mathcal{F}$ and $\mathcal{F}^{\alpha}$.

Lemma 1.8. *Let $\mathcal{F} \subseteq B^{\mathcal{X}}$ and $\alpha > 0$. Then*

$$Sdim(\mathcal{F}^{\alpha}) \leqslant fat_{\alpha/2}(\mathcal{F}), \tag{1.33}$$

$$\mathcal{M}(\alpha, \mathcal{F}, X) \leqslant \mathcal{M}(2, \mathcal{F}^{\alpha/2}, X) \tag{1.34}$$

The proof is offered to the reader as a problem.

The following lemma is the main technical part the proof of Theorem 1.10.

Lemma 1.9. *Let $|\mathcal{X}| = n$ and $B = \{0, 1, \ldots, b\}$. Let also, $\mathcal{F} \subseteq B^{\mathcal{X}}$ and $d = Sdim(\mathcal{F})$. Then*

$$\mathcal{M}_l(2, \mathcal{F}) \leqslant 2(n(b+1)^2)^{\lceil \log y \rceil},$$

where $y = \sum_{i=1}^{d} \binom{n}{i} b^i$.

Proof. Assume that $b \geqslant 2$ and define a function $t(h, n)$ as follows. Consider all pairwise separated subclasses F of the class $\mathcal{F}$ of cardinality h. Let S_h be the set of all such subclasses F. Any class $F \in S_h$ can strongly shatter some sets $X \subseteq \mathcal{X}$ with respect to some sequence s. Let k_F be the total number of all such pairs (X, s). Define $t(h, n) = \min_{F \in \mathcal{S}_h} k_F$. More formally,

$$\begin{aligned} t(h, n) = \max\{k : \forall F \subseteq \mathcal{F}, |F| = n, F \text{ pairwise separated} \\ \Rightarrow F \text{ strongly shatters at least } k \ (X, s) \text{ pairs}\}. \end{aligned}$$

When we say that F strongly shatters a pair (X, s), we mean that F strongly shatters X according to s.

Lemma 1.10. *If $t(h, n) > y$ and $Sdim(\mathcal{F}) \leqslant d$ then $\mathcal{M}_l(2, \mathcal{F}) < h$, where* $y = \sum_{i=0}^{d} \binom{n}{i} b^i$.

Proof. Assume that $\mathcal{M}_l(2, \mathcal{F}) \geqslant h$. This means that a pairwise separated set $F \subseteq \mathcal{F}$ of size $\geqslant h$ exists. Since $t(h, n) \geqslant y$, F strongly shatters at least y pairs (X, s).

Since $Sdim(\mathcal{F}) \leqslant d$, if F strongly shatters the pair (X, s) then $|X| \leqslant d$. A subset X of size i can be chosen by $\binom{n}{i}$ ways; besides,

there are $< b^i$ possible sequences s of length i (since X is strongly separated s cannot contain 0 and b). Therefore, F strongly separates less than

$$\sum_{i=1}^{d} \binom{n}{i} b^i = y$$

pairs (X, s). This contradiction proves the lemma. $\triangle$

It follows from Lemma 1.10 that in order to prove Lemma 1.9 we need to prove the inequality

$$t\left(2(n(b+1)^2)^{\lceil \log y \rceil}, n\right) \geqslant y, \tag{1.35}$$

where $y = \sum\limits_{i=1}^{d} \binom{n}{i} b^i$.

To prove the inequality (1.35) we first prove the following statement.

Lemma 1.11.

$$t(2, n) \geqslant 1 \text{ for } n \geqslant 1, \tag{1.36}$$

$$t(2mn(b+1)^2, n) \geqslant 2t(2m, n-1) \text{ for } n \geqslant 2, m \geqslant 1. \tag{1.37}$$

Proof. For any two separated functions f and g, $|f(x) - g(x)| \geqslant 2$ for at least one x, ie, these functions strongly shatters at least some singleton $\{x\}$. and so $t(2, n) \geqslant 1$, The inequality (1.36) is valid.

To prove (1.37) consider a set F containing at least $2mn(b+1)^2$ pairwise separable functions. If there are no such set F then $t(2mn(b+1)^2, n) = \infty$ and the inequality (1.37) is satisfied. Divide all function from F on pairs $\{f, g\}$. There are at least $mn(b+1)^2$ of such pairs.

Let P be the set of all such pairs. For any pair $\{f, g\} \in P$, let $\chi(f, g)$ be an x such that $|f(x) - g(x)| \geqslant 2$.

For any $x \in \mathcal{X}$, $i, j \in B$, and $j \geqslant i + 2$, define

$$bin(x, i, j) = \{\{f, g\} \in P : \chi(f, g) = x, \{f(x), g(x)\} = \{i, j\}\}.$$

The total number of such sets does not exceed

$$n\binom{b+2}{2} < n(b+1)^2/2.$$

Recall that by Lemma 1.9, we have $|\mathcal{X}| = n$.

Since the total nubber of all pairs is at least $mn(b+1)^2$, the numbers x^*, i^* and j^*, $j^* > i^* + 1$ exist such that

$$|bin(x^*, i^*, j^*)| \geqslant 2m.$$

Let us define two sets of functions

$$F_1 = \{f \in \cup bin(x^*, i^*, j^*) : f(x^*) = i^*\},$$
$$F_2 = \{g \in \cup bin(x^*, i^*, j^*) : g(x^*) = j^*\}.$$

Here, for any set A consisting of pairs, we denote by $\cup A$ the set of all elements of such pairs.

Clearly, $|F_1| = |F_2| \geqslant 2m$. The class F of functions is pairwise separated if we restrict all these functions on the set $\mathcal{X} \setminus \{x^*\}$. Indeed, the class of F, and thus the class F_1 are pairwise separated on the set $\mathcal{X}$. Therefore, for any two functions $f, f' \in F_1$ the inequality $|f(x') - f'(x')| \geqslant 2$ is valid for some x'. Moreover, $x' \in \mathcal{X} \setminus \{x^*\}$, since $f(x^*) = f'(x^*)$.

Similarly, the class of functions F_2 is also pairwise separated on the set $\mathcal{X} \setminus \{x^*\}$.

Consequently, there exists two sets U and V of size $\geqslant t(2m, n-1)$ consisting of pairs (X, s) such that F_1 strongly shatters pairs in U and F_2 strongly shatters pairs in V. Further, $|U| \geqslant t(2m, n-1)$ and $|V| \geqslant t(2m, n-1)$.

Any pair in $U \cup V$ is obviously shattered by the class F. Let $(X, s) \in U \cap V$. Then the pair $(\{x^*\} \cup X, \lfloor \frac{i^*+j^*}{2} \rfloor, s)$ is also shattered by F. This is because any functions $f \in F_1$ and $g \in F_2$ strongly shattering X also satisfy conditions $f(x^*) = i^*$ and $g(x^*) = j^*$, and $j^* \geqslant i^* + 2$. Then $g(x^*) = j^* \geqslant \frac{i^*+j^*}{2} + 1$ and $f(x^*) = i^* \leqslant \frac{i^*+j^*}{2} - 1$. Hence, at least one such function strongly shatters this set.

Indeed, let $E \subseteq X$ and $f(x) \geqslant s_i + 1$ if $x \in E$ and $f(x) \leqslant s_i - 1$ if $x \notin E$ for some sequence $s = (s_1, \ldots, s_{n-1})$. Similarly, let $g(x) \geqslant s'_i + 1$ if $x \in E$ and $f(x) \leqslant s'_i - 1$ if $x \notin E$ for some sequence $s' = (s'_1, \ldots, s'_{n-1})$. Also, $f \in F_1$ and $g \in F_2$. Then the function f strongly shatters E with respect to the sequence $s_1 = (\lfloor \frac{i^*+j^*}{2} \rfloor, s_1, \ldots, s_{n-1})$ if $x^* \notin E$ or the function g strongly shatters E with respect to the sequence $s_2 = (\lfloor \frac{i^*+j^*}{2} \rfloor, s'_1, \ldots, s'_{n-1})$ if $x^* \in E$.

Thus, the class F strongly shatters

$$|U \cup V| + |U \cap V| = |U| + |V| \geqslant 2t(2m, n-1)$$

pairs (X, s).

The inequality (1.37) and Lemma 1.11 are proved. $\triangle$

We now turn to the proof of Lemma 1.9. Applying the inequalities (1.36) and (1.37) recursively, we obtain

$$t(2(n(b+1)^2)^r, n) \geqslant 2^r t(2, n-r) \geqslant 2^r \tag{1.38}$$

for $n > r \geqslant 1$.

If $\lceil \log y \rceil < n$ then taking $r = \lceil \log y \rceil$ in (1.38) we obtain the inequality (1.9).

If $\lceil \log y \rceil \geqslant n$ then the number

$$2(n(b+1)^2)^{\lceil \log y \rceil} > (b+1)^n$$

exceeds the total number of all functions with range in B and with domain $\mathcal{X}$, $|\mathcal{X}| = n$.

Thus, a pairwise separated set F of size $2(n(b+1)^2)^{\lceil \log y \rceil}$ does not exist and hence

$$t(2(n(b+1)^2)^{\lceil \log y \rceil}, n) = \infty.$$

Lemma (1.9) is proved. $\triangle$

We can now state and prove the main result of this section – Alon, Ben-Dawid, Cesa-Bianchi and Haussler theorem [1].

Theorem 1.10. *Let $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ and $\alpha \in [0,1]$. Denote $d = \mathrm{fat}_{\alpha/4}(\mathcal{F})$. Then*

$$\mathcal{N}(\alpha, \mathcal{F}, n) \leqslant 2\left(n\left(\frac{2}{\alpha}+1\right)^2\right)^{\lceil d \log(\frac{2en}{d\alpha}) \rceil}.$$

Proof. Using the fact that the covering number does not exceed the packing number, the inequality (1.34) of Lemma 1.8, and

Lemma 1.9, we obtain the following chain of inequalities:

$$\begin{aligned}\mathcal{N}(\alpha,\mathcal{F},n) = \sup_{|X|=n} \mathcal{N}(\alpha,\mathcal{F},X) \leqslant \\ \leqslant \sup_{|X|=n} \mathcal{M}(\alpha,\mathcal{F},X) \leqslant \\ \leqslant \sup_{|X|=n} \mathcal{M}(2,\mathcal{F}^{\alpha/2},X) = \mathcal{M}(2,\mathcal{F}^{\alpha/2}) \leqslant \\ \leqslant 2(n(b+1)^2)^{\lceil \log y\rceil},\end{aligned}$$

where $b = \lceil \frac{2}{\alpha} \rceil$, $y = \sum\limits_{i=1}^{d'} \binom{n}{i} b^i$, $d' = Sdim(\mathcal{F}^{\alpha/2})$.

Note that the class $\mathcal{F}^{\alpha/2}$ satisfies the assumption of Lemma 1.9 for $b = \lceil \frac{2}{\alpha} \rceil$.

From the inequality (1.33) of Lemma 1.8, the inequality $d' \leqslant fat_{\alpha/4}(\mathcal{F}) = d$ follows. Hence,

$$y \leqslant \sum_{i=1}^{d} \binom{n}{i} b^i \leqslant b^d \sum_{i=1}^{d} \binom{n}{i} \leqslant b^d \left(\frac{en}{d}\right)^d.$$

In particular, $\log y \leqslant d \log\left(\frac{ben}{d}\right)$.

The theorem is proved. $\triangle$

Theorem 1.7 in Section 1.3 is a reformulation of this theorem with a little attenuation of estimates.

1.4. Rademacher averages

In this section, we consider another definition of the capacity of a class of functions – Rademacher averages (the sources for this section are Bousquet et al. [7], Bartlett and Mendelson [3], and Kakade and Tewari [18]). This concept allows us to obtain new upper bounds for the generalization error.

Let $z^l = (z_1, \ \ldots, \ z_l)$ be a sample of unlabeled examples, whose elements belong to some set $\mathcal{X}$ with a structure of probability space, P be a probability distribution on $\mathcal{X}$. Assume that elements of z^l are generated i.i.d. according to the probability distribution P.

Let also, $\mathcal{F}$ be a class of uniformly bounded functions defined on $\mathcal{X}$.

Let $\sigma_1,\ \ldots,\ \sigma_l$ be i.i.d. Bernoulli variables taking values $+1$ and -1 with equal probabilities: $B_{1/2}(\sigma_i = 1) = B_{1/2}(\sigma_i = -1) = 1/2$ for all $1 \leqslant i \leqslant l$. These variables are called *Rademacher variables*.

Denote $\sigma = B_{1/2}^l$ the probability of generating a sequence $\sigma_1, \ldots, \sigma_l$ of length l.

Define *the empirical Rademacher average* of the class $\mathcal{F}$ as the random variable (that is a function of random variables $z_1,\ \ldots,\ z_l$)

$$\tilde{\mathcal{R}}_l(\mathcal{F}) = E_\sigma \left(\sup_{f \in \mathcal{F}} \frac{1}{l} \sum_{i=1}^{l} \sigma_i f(z_i) \right).$$

Note that the probability distribution P on $\mathcal{X}$ generates the product probability distribution P^l on the set of all samples $z^l = (z_1,\ \ldots,\ z_l)$ of length l.

The Rademacher average of the class $\mathcal{F}$ is defined as

$$\mathcal{R}_l(\mathcal{F}) = E_{P^l}(\tilde{\mathcal{R}}_l(\mathcal{F})) = E_{P^l} E_\sigma \left(\sup_{f \in \mathcal{F}} \frac{1}{l} \sum_{i=1}^{l} \sigma_i f(z_i) \right).$$

By definition the Rademacher average is the mathematical expectation of the empirical Rademacher average with respect to probability distribution P^l.

Rademacher averages give us a powerful tool to obtain uniform convergence bounds. We present some properties of Rademacher averages, which will be used for obtaining in Section 2.7 the uniform upper bounds of generalization error.

Assume that the elements of a sample $z^l = (z_1,\ \ldots,\ z_l)$ be generated i.i.d. by some probability distribution P. By definition the empirical mean of a function f on the sample z^l equals

$$\hat{E}_{z^l}(f) = \frac{1}{l} \sum_{i=1}^{l} f(z_i).$$

The true mathematical expectation of the function f is equal to $E_P(f) = \int f(z) dP$.

In the following theorem a bound of the difference between the empirical and true expectations is presented. This bound is uniform over the functional class $\mathcal{F}$.

Theorem 1.11. *The following inequality is valid:*

$$E_{z^l\sim P^l}(\sup_{f\in\mathcal{F}}(E_P(f)-\tilde{E}(f(z^l))))\leqslant 2\mathcal{R}_l(\mathcal{F}). \tag{1.39}$$

Proof. Given a random sample $z^l=(z_1,\ \ldots,\ z_l)$, let $\tilde{z}^l=(\tilde{z}_1,\ \ldots,\ \tilde{z}_l)$ be a "ghost sample". This means that random variables $\tilde{z}_i$ are independent of each other and of z_i, $i=1,\ldots,l$, and have the same distribution as the latter.

The following chain of equalities and inequalities is valid:

$$\begin{aligned}
&E_{z^l\sim P^l}\left(\sup_{f\in\mathcal{F}}\left(E_P(f(z))-\frac{1}{l}\sum_{i=1}^{l}f(z_i)\right)\right)=\\
&=E_{z^l\sim P^l}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}E_{\tilde{z}_i\sim P}(f(\tilde{z}_i))-f(z_i)\right)\right)\leqslant\\
&\leqslant E_{z^l\sim P^l}\left(E_{\tilde{z}^l\sim P^l}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}(f(\tilde{z}_i))-f(z_i)\right)\right)\right)=\\
&=E_{z^l\tilde{z}^l\sim P^{2l}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}(f(\tilde{z}_i))-f(z_i)\right)\right)=\\
&=E_{z^l\tilde{z}^l\sim P^{2l}}E_{\sigma\sim B_{1/2}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(\tilde{z}_i))-f(z_i)\right)\right)\leqslant\\
&=E_{\tilde{z}^l\sim P^l}E_{\sigma\sim B_{1/2}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(\tilde{z}_i)\right)\right)+\\
&+E_{z^l\sim P^l}E_{\sigma\sim B_{1/2}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(z_i)\right)\right)=\\
&=2\mathcal{R}_l(\mathcal{F}).
\end{aligned} \tag{1.40}$$

The transition from the 2th line to 3th one is valid, since supremum of a sum does not exceed the sum of suprema. Inserting σ_i in the

5th line do not change the supremum, since the mathematical expectation of supremum is invariant under transposition of any variables z_i and $\tilde{z}_i$. By this reason we can insert in the 6th line the symbol of mathematical expectation $E_{\sigma\sim B_{1/2}}$.

The inequality (1.39) is proved. $\triangle$

Now we give two corollaries of Theorem 1.11.

First, the inequality (1.39) can be inverted:

Corollary 1.5. *For any function $f \in \mathcal{F}$,*

$$E_{P^l}(\sup_{f\in\mathcal{F}}(\tilde{E}_{z^l}(f) - E_P(f)) \leqslant 2\mathcal{R}_l(\mathcal{F}). \tag{1.41}$$

The inequality (1.41) follows directly from the inequality (1.39) and from the obvious equality $\mathcal{R}_l(\mathcal{F}) = \mathcal{R}_l(-\mathcal{F})$, where $-\mathcal{F} = \{-f : f \in \mathcal{F}\}$.

To prove the second corollary, we need the following lemma which is presented without a proof.

Lemma 1.12. *Let $f : \mathcal{X}^l \to \mathcal{R}$ be a function satisfying*

$$|f(z_1, \ \ldots, \ z_{i-1}, z_i, z_{i+1}, \ \ldots, \ z_l) - \\ -f(z_1, \ \ldots, \ z_{i-1}, z_i', z_{i+1}, \ \ldots, \ z_l)| \leqslant c_i$$

for all i and for all $z_1, \ \ldots, \ z_l, z_i' \in \mathcal{Z}$, where $c_1, \ \ldots, \ c_l$ are some constants.

Let also, $\tilde{z}_1, \ \ldots, \ \tilde{z}_l$ be i.i.d. random variables with range in $\mathcal{X}$ distributed according to a probability distribution P. Then

$$P^l\{f(\tilde{z}_1, \ \ldots, \ \tilde{z}_l) - E_{P^l}(f(\tilde{z}_1, \ \ldots, \ \tilde{z}_l)) \geqslant t\} \leqslant \\ \leqslant \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{l} c_i^2}\right), \tag{1.42}$$

where E_{P^l} is a symbol of the mathematical expectation with respect to the probability distribution P^l on samples of length l.

The proof of this lemma can be found in [25] and [28].

Since the condition of this lemma is satisfied, where we replace f on $-f$, the following inequality also holds:

$$P^l\{E_{P^l}(f(z_1,\ \ldots,\ z_l)) - f(z_1,\ \ldots,\ z_l) \geqslant t\} \leqslant$$
$$\leqslant \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{l} c_i^2}\right). \qquad (1.43)$$

The following corollary presents the uniform bound of the difference between the expectation of the function and the sample mean of this function:

Corollary 1.6. *Assume that all functions from a class $\mathcal{F}$ take values in $[0,1]$. Then for any $f \in \mathcal{F}$ and any $\delta > 0$, with probability $1-\delta$:*

$$E_P(f(z)) \leqslant \hat{E}_{z^l}(f) + 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2l}} \leqslant$$
$$\leqslant E_P(f(z)) \leqslant \hat{E}_{z^l}(f) + 2\tilde{\mathcal{R}}_l(\mathcal{F}) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2l}}. \qquad (1.44)$$

Proof. For any function f, the following inequality obviously holds:

$$E_P(f(z)) \leqslant \hat{E}_{z^l}(f) + \sup_{h\in\mathcal{F}}(E_P(h) - \tilde{E}_{z^l}(h)). \qquad (1.45)$$

Let us apply the inequality (1.42) of Lemma 1.12 to the second term of (1.45).

Since the function f is nonnegative and bounded by 1, one can take $c_i = 1/l$ for all $1 \leqslant i \leqslant l$. Substituting these values in the right-hand side of the inequality (1.42) and equate it to $\delta/2$, we obtain

$$\exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{l} c_i^2}\right) = e^{-2t^2 l} = \delta/2.$$

Then $t = \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}$. By (1.42), with probability $1 - \delta/2$,

$$\sup_{h \in \mathcal{F}}(E_P(h) - \tilde{E}(h)) \leqslant E_{P^l}(\sup_{h \in \mathcal{F}}(E_P(h) - \tilde{E}(h))) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \quad (1.46)$$

The inequality (1.39) says that

$$E_{P^l}(\sup_{f \in \mathcal{F}}(E_P(f) - \tilde{E}_{z^l}(f)) \leqslant 2\mathcal{R}_l(\mathcal{F}).$$

From this and from (1.46), we obtain

$$\sup_{h \in \mathcal{F}}(E_P(h) - \tilde{E}_{z^l}(h)) \leqslant 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \quad (1.47)$$

Hence, with probability $1 - \delta/2$, the bound

$$E_P(f) \leqslant \tilde{E}(f)) + 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}} \quad (1.48)$$

is valid for all $f \in \mathcal{F}$. Thus, the inequality (1.44) is proved.

Similarly, using the inequality (1.43) of Lemma 1.12, we obtain that, with probability $1 - \delta/2$,

$$\mathcal{R}_l(\mathcal{F}) \leqslant \tilde{\mathcal{R}}_l(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \quad (1.49)$$

By (1.48) and (1.49) we obtain that, with probability $1 - \delta$, the inequality (1.44) holds. $\triangle$

In order to use our Rademacher bound, we need to find Rademacher complexities of composition $\phi \circ \mathcal{F} = \phi(\mathcal{F}) = \{\phi(f) : f \in \mathcal{F}\}$, where ϕ is some function.

Theorem 1.12. *Assume that ϕ be L-Lipschitz continuous function, ie*

$$|\phi(x) - \phi(y)| \leqslant L|x - y|$$

for all x and y. Then

$$\tilde{\mathcal{R}}_l(\phi(\mathcal{F})) \leqslant L\tilde{\mathcal{R}}_l(\mathcal{F}), \quad (1.50)$$

$$\mathcal{R}_l(\phi(\mathcal{F})) \leqslant L\mathcal{R}_l(\mathcal{F}). \quad (1.51)$$

Proof. Let $z^l = (z_1, \ \ldots, \ z_l)$ be a random sample distributed according to a probability distribution P, $\sigma_1, \ \ldots, \ \sigma_l$ be the i.i.d. Bernoulli random variables taking values in the set $\{-1, +1\}$, and let σ be the probability distribution on the set of all such sequences of length l induced by P.

The transformations given below are valid for mathematical expectations $E = E_\sigma$ and $E = E_{P^l} E_\sigma$ simultaneously. Thus we will prove both inequalities (1.50) and (1.51) simultaneously.

By definition the empirical Rademacher average of the class $\phi(\mathcal{F})$ is equal to

$$\mathcal{R}_l(\phi(\mathcal{F})) = E\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi(f(z_i))\right). \tag{1.52}$$

For simplicity, we assume that $L = 1$. [9] We need to prove that

$$\mathcal{R}_l(\phi(\mathcal{F})) \leqslant \mathcal{R}_l(\mathcal{F}) = E\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(z_i)\right). \tag{1.53}$$

We make the transition from (1.52) to (1.53) step-by-step. At each step, we consider a sequence of auxiliary functions $(\phi_1, \ \ldots, \ \phi_l)$, where each function ϕ_i is ϕ or it is identity function I.

At the first step all the functions are ϕ: $\phi_i = \phi$ for all i, at the last step of all these functions are identity functions: $\phi_i = I$ for all i.

We also assume that at each step, except the last one, $\phi_1 = \phi$. In the transition to the next step the next function $\phi_i = \phi$ will be replaced by the identity function: $\phi_i' = I$. This will be achieved by

[9] One can replace the function ϕ on ϕ/L.

the following chain of equalities and inequalities:

$$
\begin{aligned}
E(\sup_{f\in\mathcal{F}}\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi_i(f(z_i))) = \\
\frac{1}{2l}E(\sup_{f\in\mathcal{F}}(\phi(f(z_1))+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)))+ \\
+\sup_{f\in\mathcal{F}}(-\phi(f(z_1))+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)))) = \\
=\frac{1}{2l}E(\sup_{f,f'\in\mathcal{F}}(\phi(f(z_1))+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))- \\
-\phi(f'(z_1))+\sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i)))) \leqslant \\
\leqslant\frac{1}{2l}E(\sup_{f,f'\in\mathcal{F}}(|f(z_1)-f'(z_1)|+ \\
+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))+\sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i)))) = \\
=\frac{1}{2l}E(\sup_{f,f'\in\mathcal{F}}(f(z_1)-f'(z_1)+ \\
+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))+\sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i)))) \leqslant \\
\leqslant\frac{1}{2l}E(\sup_{f\in\mathcal{F}}(f(z_1)+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))+ \\
\sup_{f'\in\mathcal{F}}(-f'(z_1)+\sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i))) = \\
=E(\sup_{f\in\mathcal{F}}\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi_i'(f(z_i))), \qquad (1.54)
\end{aligned}
$$

where the collection of functions $\phi_1', \ \ldots, \ \phi_l'$ contains one more identity function than the previous collection $\phi_1, \ \ldots, \ \phi_l$.

In transition from the 1th line to the 2th and 3th lines, we take the mathematical expectation by σ_1; after that one can still consider E as the expectation by the whole set σ, because now the variable σ_1 absent.

In transition from the 4th and 5th lines to the 6th and 7th, we have used an observation that the supremum is achieved by non-negative values of the difference $\phi(f(z_1))-\phi(f'(z_1))$, so we can replace it by its absolute value. After that, Lipschitz's condition has used for $L = 1$. A similar reason was used in transition from the 6th and 7th lines to the 8th and 9th lines.

Transition from the 8th and 9th lines to the 10th line has done by the same reason as transition from the 1th line to the 2th and 3th lines.

Applying several times the chain of transformations (1.54) we obtain the expression

$$E\left(\sup_{f\in\mathcal{F}}\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi_i'(f(z_i))\right), \tag{1.55}$$

where all ϕ_i' are identity functions, and so, the sum (1.55) is equal to $\mathcal{R}_l(\mathcal{F})$.

The first line of the chain (1.54) is equal to $\mathcal{R}_l(\phi(\mathcal{F}))$ for $E = E_{P^l}E\sigma$ or to $\tilde{\mathcal{R}}_l(\phi(\mathcal{F}))$ for $E = E_\sigma$.

Thus, the inequalities (1.50) and (1.51) are satisfied, and the theorem is proved. $\triangle$

1.5. Rademacher averages and other capacity measures

In this section we study connection of Rademacher average with other known measures of capacity of the classes of functions – the growth function $B_{\mathcal{F}}(l)$ and the covering number $\mathcal{N}(\alpha, \mathcal{F}, l)$.

Comparison with the growth function

We need the following auxiliary statement - Massar Lemma:

Lemma 1.13. *Let A be a finite subset of $\mathcal{R}^l$ and $\sigma_1, \ \ldots, \ \sigma_l$ be i.i.d. random variables. Then*

$$E_\sigma \left(\sup_{a \in A} \frac{1}{m} \sum_{i=1}^{l} \sigma_i a_i \right) \leqslant \sup_{a \in A} \|a\| \frac{\sqrt{2 \ln |A|}}{l},$$

where $a = (a_1, \ \ldots, \ a_l)$.

Proof. Denote $E = E_\sigma$. The following chain of equalities and inequalities is valid:

$$\begin{aligned}
&\exp\left(\lambda E\left(\sup_{a\in A}\sum_{i=1}^{l}\sigma_i a_i\right)\right)\leqslant\\
&\leqslant E\left(\exp\left(\lambda\sup_{a\in A}\sum_{i=1}^{l}\sigma_i a_i\right)\right)=\\
&=E\left(\sup_{a\in A}\exp\left(\lambda\sum_{i=1}^{l}\sigma_i a_i\right)\right)\leqslant\\
&\leqslant E\left(\sum_{a\in A}\exp\left(\lambda\sum_{i=1}^{l}\sigma_i a_i\right)\right)=\\
&=\sum_{a\in A}E\left(\exp\left(\lambda\sum_{i=1}^{l}\sigma_i a_i\right)\right)=\\
&=\sum_{a\in A}\prod_{i=1}^{l}E(\exp(\lambda\sigma_i a_i))=\\
&=\sum_{a\in A}\prod_{i=1}^{l}\frac{e^{\lambda a_i}+e^{-\lambda a_i}}{2}\leqslant\\
&\leqslant\sum_{a\in A}\prod_{i=1}^{l}e^{\lambda^2\|a\|^2/2}\leqslant\\
&\leqslant|A|e^{\lambda^2 r^2/2},
\end{aligned}$$

where $r = \sup_{a \in A} \|a\|$. Here, in the transition from the first to the second line, the convexity of the exponent was used. In the transition from

the 7th row to the 8th one, we use the inequality $e^x + e^{-x} \leqslant 2e^{x^2/2}$. Other transitions are obvious.

Taking logarithm of the first and the second lines of this inequality, we obtain the inequality:

$$E\left(\sup_{a\in A}\sum_{i=1}^{l}\sigma_i a_i\right) \leqslant \frac{\ln|A|}{\lambda} + \frac{\lambda r^2}{2}. \tag{1.56}$$

It is easy to verify that the right-hand side of (1.56) attains its minimum at $\lambda = \sqrt{2\ln|A|/r^2}$. Substituting this value in the right-hand side of (1.56), we obtain:

$$E\left(\sup_{a\in A}\sum_{i=1}^{l}\sigma_i a_i\right) \leqslant r\sqrt{2\ln|A|}.$$

Lemma is proved. $\triangle$

The inequality between the Rademacher average and the growth function is presented in the following theorem.

Theorem 1.13. *Let $\mathcal{F}$ be a class of indicator functions taking values in the set $\{-1, +1\}$. Then*

$$\mathcal{R}_l(\mathcal{F}) \leqslant \sqrt{\frac{2\ln B_{\mathcal{F}}(l)}{m}}$$

for all l.

Proof. Let $E = E_{P^l}$ and a binary string $a = (a_1, \ldots, a_l)$ represents all values $(f(z_1), \ldots, f(z_l))$. The following chain of inequalities is valid:

$$\begin{aligned}
\mathcal{R}_l(\mathcal{F}) = EE_\sigma\left(\sup_a \frac{1}{l}\sum_{i=1}^{l}\sigma_i a_i\right) \leqslant \\
\leqslant E\left(\sup_a \|a\|\frac{\sqrt{2\ln|\mathcal{F}_{|X^l|}}}{l}\right) \leqslant \\
\leqslant E\left(\sqrt{l}\frac{\sqrt{2\ln B_{\mathcal{F}}(l)}}{l}\right) = \\
= \sqrt{\frac{2\ln B_{\mathcal{F}}}{l}}.
\end{aligned}$$

In transition from the 1th line to 2th, Lemma 1.13 has used, in transition from 2th line to 3th, we have used the value of the norm of the binary vector $\|a\| = \sqrt{l}$. We have used also the definition of the growth function.

Theorem is proved. $\triangle$

Comparison with the covering number

Let a set $\mathcal{X}$ is equipped by some probability distribution P and $x^l = (x_1, \ \ldots, \ x_l)$ be a random sample in $\mathcal{X}$. Let also, $\mathcal{F}$ be a class of uniformly bounded functions defined on $\mathcal{X}$ with the range in $[-1, 1]$.

Recall the norm $l_{x^l}(f, g) = \sup_{1 \leqslant i \leqslant l} |f(x_i) - g(x_i)|$ on $\mathcal{F}$ and the corresponding covering number $\mathcal{N}(\alpha, \mathcal{F}, x^l)$ associated with the sample x^l. The covering number is equal to the minimal size of sets $B \subseteq \mathcal{F}$ such that for any $f \in \mathcal{F}$ an $g \in B$ exists for that $l_{x^l}(f, g) < \alpha$.

Theorem 1.14. *The empirical Rademacher average satisfies:*

$$\tilde{\mathcal{R}}_l(\mathcal{F}) \leqslant \inf_{\alpha} \left(\sqrt{\frac{2 \ln \mathcal{N}(\alpha, \mathcal{F}, x^l)}{l}} + \alpha \right). \tag{1.57}$$

Proof. Let B be a minimal cover of the class $\mathcal{F}$ with respect to the sample x^l.

We can assume that the domain of functions from B is $\{x_1, \ \ldots, \ x_l\}$. Let also,

$$B_\alpha(g) = \{f \in \mathcal{F} : l_{x^l}(f, g) < \alpha\}.$$

By definition of the cover $\cup_{g\in B} B_\alpha(g) = \mathcal{F}$. Then

$$\tilde{\mathcal{R}}_l(\mathcal{F}) = E_\sigma\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(x_i)\right)\right) =$$

$$= E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(x_i)\right)\right) =$$

$$= E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i g(x_i) + \frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(x_i)-g(x_i))\right)\right) \leqslant$$

$$\leqslant E_\sigma\left(\sup_{g\in B}\frac{1}{l}\sum_{i=1}^{l}\sigma_i g(x_i)\right) +$$

$$+E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(x_i)-g(x_i))\right). \quad (1.58)$$

For mathematical expectation from the last line of (1.58), the following inequality holds:

$$E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(x_i)-g(x_i))\right) =$$

$$= E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\left|\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(x_i)-g(x_i))\right|\right) \leqslant$$

$$\leqslant E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\frac{1}{l}\sum_{i=1}^{l}\sigma_i|f(x_i)-g(x_i)|\right) \leqslant \alpha. \quad (1.59)$$

By Lemma 1.13:

$$E_\sigma\left(\sup_{g\in B}\frac{1}{l}\sum_{i=1}^{l}\sigma_i g(x_i)\right) \leqslant$$

$$\leqslant \sup_{g\in B}\|g\|\frac{\sqrt{2\ln|B|}}{l} \leqslant$$

$$\leqslant \sqrt{\frac{2\ln|B|}{l}} = \sqrt{\frac{2\ln\mathcal{N}(\alpha,\mathcal{F},x^l)}{l}}. \quad (1.60)$$

We have used $\|g\| = \sqrt{\sum_{i=1}^{l} g^2(x_i)} \leqslant \sqrt{l}$, since the size of domain of g is equal to l and this function is bounded by one.

Combining the inequalities (1.59) and (1.60), we obtain:

$$\tilde{\mathcal{R}}_l(\mathcal{F}) \leqslant \left(\sqrt{\frac{2 \ln \mathcal{N}(\alpha, \mathcal{F}, x^l)}{l}} + \alpha \right). \tag{1.61}$$

Since the inequality (1.61) is valid for all $\alpha > 0$, its is valid for infimum by $\alpha > 0$. From this the inequality (1.57) follows. Theorem is proved. $\triangle$

Theorem 1.14 clearly implies a similar inequality between Rademacher average and covering number.

Corollary 1.7.

$$\mathcal{R}_l(\mathcal{F}) \leqslant \inf_{\alpha} \left(\sqrt{\frac{2 \ln \mathcal{N}(\alpha, \mathcal{F}, l)}{l}} + \alpha \right).$$

For more information on Rademacher averages see Bartlett et al. [3], Bartlett et al. [4], Ledoux and Talagrand [21].

1.6. Problems

1. Give a complete proof of Lemmas 1.3 and 1.4.

2. Let Z be an infinite set and

$$\mathcal{P}_k(Z) = \{A : A \subseteq Z \& |A| \leqslant k\}$$

be a set of all its subsets containing no more than k elements, f_A be the characteristic function of a subset A, ie, a function equal to 1 on A and 0 on its complement. Let also, $\mathcal{H}_Z$ be a class of all characteristic functions. Prove that the growth function $B_{\mathcal{H}_Z}(l)$ satisfies

$$B_{\mathcal{H}_Z}(l) = 2^l$$

for $l \leqslant k$, and

$$B_{\mathcal{H}_Z}(l) = \sum_{i=0}^{k} \binom{l}{i}$$

for $l > k$.

3. Compute the values of growth function: $B_H(3)$, $B_H(4)$, $B_H(5)$, ..., where
a) H be a class of all homogeneous linear classifiers;
b) H be a class of all linear classifiers;
c) H be a class of all classifiers defining by polynomials of 2th degree, 3th degree, and so on.

4. Give examples of classes of indicator functions, for which VC-dimension is equal to ∞ (*Note*: Consider the class of functions $\mathcal{F} = \{\text{sign}(\sin(tx)) : t \in \mathcal{R}\}$. For any l, let $x_i = 2\pi 10^{-i}$, $i = 1, \ldots, l$, and $\delta_1, \ldots, \delta_l$ be an arbitrary set or real numbers from $\{0, 1\}$ representing the partition of elements x_i on two classes. Prove that for $t = \frac{1}{2}\left(\sum_{i=1}^{l}(1-\delta_i)10^i + 1\right)$ the equalities $\text{sign}(\sin(tx_i)) = \delta_i$ hold for all i, where $sig(r) = 1$ for $r \geqslant 0$ and $sig(r) = 0$ for $r < 0$ (see [34])).

5. Check VC-dimension of the class $\mathcal{F}$ of classifiers:

a) VC-dimension of the class $\mathcal{F}$ defined by convex polygons in $\mathcal{R}^2$ is ∞;

b) VC-dimension of the class $\mathcal{F}$ defined by axis-aligned rectangles in $\mathcal{R}^2$ is 4;

c) Find VC-dimension of the class $\mathcal{F}$ defined by convex polygons with d vertices in $\mathcal{R}^2$.

6. Obtain a bound 3) of Theorem 1.5 for the class of all classifiers defined by linear functions.

7. Prove that the recurrence relation (1.26) has the solution

$$\Phi(n, l) = \begin{cases} 2^l \text{ if } l \leqslant n \\ 2\sum\limits_{i=1}^{n-1} \binom{l-1}{i} \text{ if } l > n. \end{cases}$$

8. Let $\mathcal{G}$ be a k-dimensional vector space of functions on $\mathcal{R}^n$ and

$$\mathcal{F} = \{f : f(\bar{x}) = \text{sign}(g(\bar{x})) : g \in \mathcal{G}\}.$$

Prove that VC-dimension of class $\mathcal{F}$ is less or equal to k (*Hint*: The class of all homogeneous linear functions is an example of such functional space $\mathcal{G}$).

9. Let $\mathcal{L}$ be the class of linear functions $f(\bar{x}) = (\bar{w} \cdot \bar{x})$, where $\|\bar{w}\|_2 = \sqrt{(\bar{w} \cdot \bar{w})} \leqslant A$ and $\|\bar{x}\|_2 \leqslant R$. Prove that the Rademacher complexity is bounded:

$$\mathcal{R}_l(\mathcal{L}) \leqslant \frac{AR}{\sqrt{l}}.$$

Chapter 2

Support vector machines

The problem of classification and regression using Support Vector Machines (SVM) aims to develop efficient algorithmic techniques for constructing an optimal separating hyperplane in the feature space of high dimension. The optimality is in the sense of minimizing the upper bounds of generalization error.

2.1. Optimal hyperplane

We first consider the case of a fully separated training sample, ie, the case where training may be carried out without errors.

An ordered sample $S = ((\bar{x}_1, y_1), (\bar{x}_2, y_2), \ \ldots, \ (\bar{x}_l, y_l))$, where $\bar{x}_i \in \mathcal{R}^n$ and $y_i \in \{-1, 1\}$, $i = 1, \ \ldots, \ l$, is called *separable* by a hyperplane $(\bar{w} \cdot \bar{x}_i) - c = 0$ if there exists a vector $\bar{w}$ of length $(|\bar{w}| = 1)$ and the number c such that

$$\begin{aligned} (\bar{w} \cdot \bar{x}_i) - c > 0 \text{ if } y_i = 1, \\ (\bar{w} \cdot \bar{x}_i) - c < 0 \text{ if } y_i = -1. \end{aligned} \tag{2.1}$$

In the case, where a separating hyperplane $(\bar{w} \cdot \bar{x}_i) - c = 0$ exists, define

$$\begin{aligned} c_1(\bar{w}) = \min_{y_i=1} (\bar{w} \cdot \bar{x}_i), \\ c_2(\bar{w}) = \max_{y_i=-1} (\bar{w} \cdot \bar{x}_i). \end{aligned} \tag{2.2}$$

By definition $c_1(\bar{w}) > c_2(\bar{w})$. Besides, $c_1(\bar{w}) > c > c_2(\bar{w})$ if the hyperplane $(\bar{w} \cdot \bar{x}_i) - c = 0$ fully separates the sample.

Define

$$\rho(\bar{w}) = \frac{c_1(\bar{w}) - c_2(\bar{w})}{2}. \tag{2.3}$$

Then $\rho(\bar{w}) = \frac{1}{2}((c_1(\bar{w}) - c) + (c - c_2(\bar{w}))$ is equal to the half of the sum of the distances from the nearest points of the top and bottom to the separating hyperplane $(\bar{w} \cdot \bar{x}) - c = 0$ (see (2.1)).

Assume that a sample S is separable, ie, a number c exists such that (2.1) satisfies.

The maximum of the continuous function $\rho(\bar{w})$ defined on the compact set $\{\bar{w} : |\bar{w}| \leqslant 1\}$ exists. Let $\bar{w} = \bar{w}_0$ be a maximum point.

Lemma 2.1. *The hyperplane* $(\bar{w}_0 \cdot \bar{x}) - c_0 = 0$, *where* $c_0 = \frac{1}{2}(c_1(\bar{w}_0) + c_2(\bar{w}_0))$, *separates the sample* S. *It is exactly in the middle between the nearest points of the top and bottom of positive and negative parts of the sample.*

Proof. Indeed, if $y_i = 1$ then

$$(\bar{w}_0 \cdot \bar{x}_i) - c_0 \geqslant c_1(\bar{w}_0) - \frac{c_1(\bar{w}_0) + c_2(\bar{w}_0)}{2} = \\ = \frac{c_1(\bar{w}_0) - c_2(\bar{w}_0)}{2} > 0. \tag{2.4}$$

If $y_i = -1$ then

$$(\bar{w}_0 \cdot \bar{x}_i) - c_0 \leqslant c_2(\bar{w}_0) - \frac{c_1(\bar{w}_0) + c_2(\bar{w}_0)}{2} = \\ = -\frac{c_1(\bar{w}_0) - c_2(\bar{w}_0)}{2} < 0. \tag{2.5}$$

The rest part of the lemma is left to the reader as a problem. $\triangle$

The hyperplane $(\bar{w}_0 \cdot \bar{x}) - c_0 = 0$ is called *optimal*. For this hyperplane, the sum of the distances from the nearest to it (top and bottom) sample points is maximal among all separating S hyperplanes.

Lemma 2.2. *The optimal hyperplane is a unique hyperplane such that the sum of the distances from the nearest to it, from above and below, sample points is maximal among all separating S hyperplanes located at equal distances from them.*

Proof. The maximum point $\bar{w}_0$ of the continuous function $\rho(\bar{w})$ defined on the compact $\{\bar{w} : \|\bar{w}\| \leqslant 1\}$ is achived on the boundary, since otherwise the vector $\bar{w}^* = \frac{\bar{w}_0}{\|\bar{w}_0\|}$ would satisfy $\|\bar{w}^*\| = 1$ and $\rho(\bar{w}^*) = \frac{\rho(\bar{w}_0)}{\|\bar{w}_0\|} > \rho(\bar{w}_0)$.

This maximum is unique because the function $\rho(\bar{w})$ is concave (see a problem in Section 2.12).

If its maximum is attained at two different points lying on the boundary of the compact, it would also attained at an interior point, contrary to has just been proved. $\triangle$

Let us consider an equivalent definition of the optimal separating hyperplane. On the basis of this definition an algorithmically efficient method for constructing an optimal hyperplane in the form of quadratic programming problem will be developed. The exact algorithm constructed by this method will be given in the next section.

Find a vector $\bar{w}_0$ and a threshold b_0 such that

$$\begin{aligned} (\bar{w}_0 \cdot \bar{x}_i) + b_0 \geqslant 1 \quad y_i = 1, \\ (\bar{w}_0 \cdot \bar{x}_i) + b_0 \leqslant -1 \quad y_i = -1, \end{aligned} \tag{2.6}$$

where $i = 1, \ \ldots, \ l$, and such that the vector $\|\bar{w}_0\|$ has the smallest possible norm.

Theorem 2.1. *A vector $\bar{w}_0$ minimizing $\|\bar{w}\|^2$ under constraints* (2.6) *defines the optimal hyperplane with the weight vector $\bar{w}_0^* = \frac{\bar{w}_0}{\|\bar{w}_0\|}$. The margin between the optimal hyperplane and separated vectors is equal*:

$$\rho(\bar{w}_0^*) = \max_{\|\bar{w}\|=1} \left(\frac{1}{2} (\min_{y_i=1} (\bar{w} \cdot \bar{x}_i) - \max_{y_i=-1} (\bar{w} \cdot \bar{x}_i)) \right) = \frac{1}{\|\bar{w}_0\|}.$$

Proof. It holds

$$\rho(\bar{w}_0^*) = \frac{1}{2} \left(c_1 \left(\frac{\bar{w}_0}{\|\bar{w}_0\|} \right) - c_2 \left(\frac{\bar{w}_0}{\|\bar{w}_0\|} \right) \right) \geqslant \frac{1}{\|\bar{w}_0\|},$$

since by (2.6)

$$c_1\left(\frac{\bar{w}_0}{\|\bar{w}_0\|}\right) \geqslant \frac{1-b_0}{\|\bar{w}_0\|},$$
$$c_2\left(\frac{\bar{w}_0}{\|\bar{w}_0\|}\right) \leqslant \frac{-1-b_0}{\|\bar{w}_0\|}.$$

It remains to prove that the inequality $\rho(\bar{w}_0^*) > \frac{1}{\|\bar{w}_0\|}$ is impossible. Assume the contrary. Let us define the vector $\bar{w}_1 = \frac{\bar{w}_0^*}{\rho(\bar{w}_0^*)}$. We have

$$\|\bar{w}_1\| = \frac{\|\bar{w}_0^*\|}{\rho(\bar{w}_0^*)} < \|\bar{w}_0\|,$$

since $\|\bar{w}_0^*\| = 1$.

The vector $\bar{w}_1$ satisfies the constraints (2.6) for $b_0 = -\frac{c_1(\bar{w}_1)+c_2(\bar{w}_1)}{2}$. Indeed, for $y_i = 1$:

$$(\bar{w}_1 \cdot \bar{x}_i) - \frac{c_1(\bar{w}_1) + c_2(\bar{w}_1)}{2} =$$
$$= \frac{1}{\rho(w_0^*)}(\bar{w}_0^* \cdot \bar{x}_i) - \frac{c_1(\bar{w}_0^*) + c_2(\bar{w}_0^*)}{2\rho(w_0^*)} \geqslant$$
$$\geqslant \frac{c_1(\bar{w}_0^*)}{\frac{1}{2}(c_1(\bar{w}_0^*) - c_2(\bar{w}_0^*))} - \frac{c_1(\bar{w}_0^*) + c_2(\bar{w}_0^*)}{c_1(\bar{w}_0^*) - c_2(\bar{w}_0^*)} = 1.$$

The case $y_i = -1$ is considered similarly.

This is a contradiction, since the norm of the vector $\bar{w}_1^*$ is less than the norm of the vector $\bar{w}_0^*$. Therefore, $\rho(\bar{w}_0^*) = \frac{1}{\|\bar{w}_0\|}$. Theorem is proved. $\triangle$

By definition of $\bar{w}_0^*$:

$$\rho(\bar{w}_0^*) = \max_{\|\bar{w}\|=1} \rho(\bar{w}) = \frac{1}{\|\bar{w}_0\|}.$$

By Theorem 2.1 the quantity $\rho(\bar{w}_0^*) = \frac{1}{\|\bar{w}_0\|}$ is equal to the distance from the nearest points (positive or negative) of the sample to the optimal hyperplane

$$(\bar{w}_0^* \cdot \bar{x}) - \frac{c_1(\bar{w}_0^*) + c_2(\bar{w}_0^*)}{2} = 0,$$

which is located at equal distances between hyperplanes:

$$(\bar{w}_0^* \cdot \bar{x}) - \frac{c_1(\bar{w}_0^*) + c_2(\bar{w}_0^*)}{2} = \pm 1$$

optimally separating vectors of the positive and negative parts of the sample.

The equation of the optimal hyperplane can also be written as

$$(\bar{w}_0 \cdot \bar{x}) - \frac{c_1(\bar{w}_0) + c_2(\bar{w}_0)}{2} = 0.$$

2.2. Algorithm for constructing the optimal hyperplane

In this section, we present an algorithm for constructing an optimal hyperplane.

Two sets of conditions (2.6) can be written as

$$y_i((\bar{w} \cdot \bar{x}_i) + b) \geqslant 1 \tag{2.7}$$

for $i = 1, \ \ldots, \ l$.

According to the results of the previous section, to find the optimal hyperplane, we have to minimize norm of the weight vector $\|\bar{w}\|$ subject to the constraints (2.7).

The section 2.10 (below) shows that to solve the quadratic optimization problem

$$(\bar{w} \cdot \bar{w}) = \sum_{i=1}^{l} w_i^2 \to \min$$

under constraints (2.6) (or equivalent constraints (2.7)) we have to form the Lagrangian

$$L(\bar{w}, b, \bar{\alpha}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) - \sum_{i=1}^{l} \alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b) - 1), \tag{2.8}$$

where $\alpha_i \geqslant 0$ are Lagrange multipliers.

In order to find the saddle point of the Lagrangian (2.8), it is necessary to minimize it over $\bar{w}$ and b, and then, maximize it over the Lagrange multipliers under constraints $\alpha_i \geqslant 0$, $i = 1, \ \ldots, \ l$.

A necessary condition for the minimum of the Lagrangian is

$$\frac{\partial L(\bar{w}, b, \bar{\alpha})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} \alpha_i y_i \bar{x}_i = \bar{0}, \tag{2.9}$$

$$\frac{\partial L(\bar{w}, b, \bar{\alpha})}{\partial b} = \sum_{i=1}^{l} \alpha_i y_i = 0. \tag{2.10}$$

It follows from (2.9) – (2.10) that

$$\bar{w} = \sum_{i=1}^{l} \alpha_i y_i \bar{x}_i, \tag{2.11}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0. \tag{2.12}$$

Substitute (2.11) into (2.8) and denote $W(\bar{\alpha}) = L(\bar{w}, b, \bar{\alpha})$. Taking into account (2.12), one obtains

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\bar{x}_i \cdot \bar{x}_j). \tag{2.13}$$

To construct the optimal hyperplane one has to maximize the function (2.13) under constraints (2.12) and $\alpha_i \geqslant 0$, where $i = 1, \ \ldots, \ l$.

Let this maximum is attained for $\alpha_i = \alpha_i^0$, $i = 1, \ \ldots, \ l$. Then the solution of the problem of finding the optimal hyperplane has the form

$$\bar{w}_0 = \sum_{i=1}^{l} \alpha_i^0 y_i \bar{x}_i. \tag{2.14}$$

In this case,

$$b_0 = \frac{\min\limits_{y_i=1}(\bar{w}_0 \cdot \bar{x}_i) + \max\limits_{y_i=-1}(\bar{w}_0 \cdot \bar{x}_i)}{2}.$$

The optimal solution $\bar{w}_0$ and b_0 must satisfy Kuhn–Tucker conditions

$$\alpha_i^0(y_i((\bar{w}_0 \cdot \bar{x}_i) + b) - 1) = 0 \tag{2.15}$$

for $i = 1, \ \ldots, \ l$.

From this follows that $\alpha_i^0 > 0$ can be only for those i, for which $y_i((\bar{w}_0 \cdot \bar{x}_i) \ + \ b_0) - 1 = 0$. The corresponding vectors $\bar{x}_i$ lie on the hyperplanes $(\bar{w}_0 \cdot \bar{x}_i) + b_0 = \pm 1$. We call such vectors *support vectors*.

The weight vector $\bar{w}_0$ of the optimal hyperlane is expanded with nonzero coefficients on support vectors $\bar{x}_{i_s}$, $s = 1, \ \ldots, \ k$, (k is the number of support vectors):

$$\bar{w}_0 = \sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} \bar{x}_{i_s}.$$

The optimal hyperplane has the form

$$\sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} (\bar{x}_{i_s} \cdot \bar{x}) + b_0 = 0. \tag{2.16}$$

Others, non-support vectors, can not be taken into account, for example, they can be changed, with the optimal hyperplane does not change.

We also present some relations with the support vectors:

$$\|\bar{w}_0\|^2 = (\bar{w}_0 \cdot \bar{w}_0) = \sum_{s,q=1}^{k} \alpha_{i_s}^0 \alpha_{i_q}^0 y_{i_s} y_{i_q} (\bar{x}_{i_s} \cdot \bar{x}_{i_q}) \tag{2.17}$$

and

$$W(\bar{\alpha}^0) = \sum_{s,q=1}^{k} \alpha_{i_s}^0 - \frac{1}{2} \|\bar{w}_0\|^2.$$

Summing (2.15) over i, we obtain

$$\sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} (\bar{w}_0 \cdot \bar{x}_{i_s}) + b_0 \sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} = \sum_{s=1}^{k} \alpha_{i_s}^0.$$

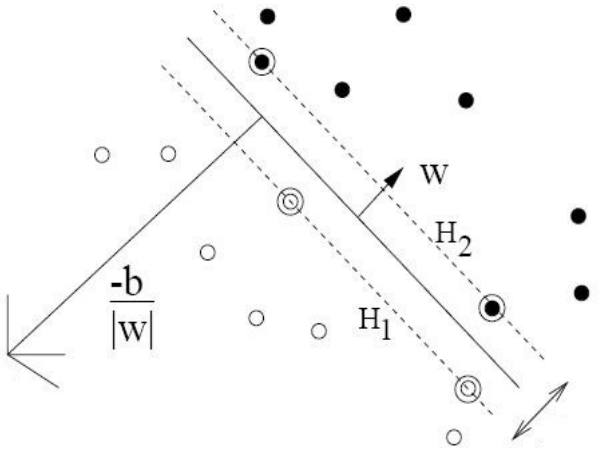


Fig. 1.1. Support vectors locate on the boundary hyperplanes H_1 and H_2

By (2.10) the second term in this sum is 0. Hence, using (2.17), we obtain

$$\sum_{s=1}^{k} \alpha_{i_s}^0 = \sum_{s,q=1}^{k} \alpha_{i_s}^0 \alpha_{i_q}^0 y_{i_s} y_{i_q} (\bar{x}_{i_s} \cdot \bar{x}_{i_q}) = \|\bar{w}_0\|^2.$$

Therefore, $W(\bar{\alpha}^0) = \frac{1}{2}\|\bar{w}_0\|^2$. We have also

$$\|\bar{w}_0\| = \frac{1}{\sqrt{\sum\limits_{s=1}^{k} \alpha_{i_s}^0}}.$$

2.3. Upper bound for generalization error in terms of support vectors

We have shown that the optimal separating hyperplane is defined not by all vectors of the sample S, but only by support vectors. One can consider the transition from a sample S to the separating hyperplane $\rho(S)$ as an information compression scheme applied to the sample S. A subsample $\hat{S}$ composed by support vectors defines the same hyperplane: $\rho(\hat{S}) = \rho(S)$.

Assume that size of $\hat{S}$ is d. There are $\binom{l}{d}$ subsets of of the sample S. Any such subset consider the corresponding classifier $h_{\hat{S}}$ defined by S. For any such classifier $h_{\hat{S}}$, the probability that it is compatible with the other $l - d$ points but has a generalization error more than ϵ, is bounded by $(1 - \epsilon)^{l-d} \leqslant \exp(-\epsilon(l - d))$.

Therefore, the probability that any classifier $h_{\hat{S}}$ defined by a subset of size d is compatible with the other $l - d$ points but has a generalization error more than ϵ, is bounded by

$$\binom{l-d}{d} \exp(-\epsilon(l-d)).$$

Thus, we have proved the theorem:

Theorem 2.2. *For any probability distribution P on $X \times \{-1, 1\}$, with P^l-probability $1 - \delta$, any classifier $\hat{S}$ defined by a subset of a random sample S of size d has a generalization error no more than*

$$\operatorname{err}_P(h_{\hat{S}}) \leqslant \frac{1}{l-d}\left(d \ln\left(\frac{el}{d}\right) + \ln\left(\frac{l}{\delta}\right)\right).$$

This theorem implies that, for $d > 2$ and sufficiently large l

$$\operatorname{err}_P(h_{\hat{S}}) \leqslant \frac{d \ln l}{l-d},$$

where d is the number of support vectors.

Unfortunately, in practical applications, the number of support vectors is often comparable in order to sample size. For this reason, the bound of Theorem 2.2 is useless.

2.4. SVM-method in feature space

SVM method is based on the following idea. Let an ordered training sample $S = ((\bar{x}_1, y_1), (\bar{x}_2, y_2), \ \ldots, \ (\bar{x}_l, y_l))$ be given.

Sample vectors $\bar{x}_1, \ldots, \bar{x}_l$ belonging to the space $\mathcal{R}^n$ are mapped to a feature space of a higher dimension $\mathcal{R}^N$ using some non-linear mapping $\bar{\phi} : \mathcal{R}^n \to \mathcal{R}^N$ chosing a priory:

$$\bar{x} = (x_1, \ \ldots, \ x_n) \to \bar{\phi}(\bar{x}) = (\phi_1(\bar{x}), \ \ldots, \ \phi_N(\bar{x})). \tag{2.18}$$

Consider the induced training sample

$$\phi(S) = ((\bar{\phi}(\bar{x}_1), y_1) \ \ldots, \ (\bar{\phi}(\bar{x}_l), y_l))$$

in the feature space $\mathcal{R}^N$.

In general case the mapping (2.18) can be non-invertible. The initial space $\mathcal{R}^n$ is mapped by $\bar{x} \to \bar{\phi}(\bar{x})$ to a subset of the features space $\mathcal{R}^N$. We construct the optimal hyperplane separating the vectors $\bar{\phi}(\bar{x}_1), \ \ldots, \ \bar{\phi}(\bar{x}_l)$ in the space $\mathcal{R}^N$.

Example. Suppose that we solve a classification problem in the n-dimensional space using polynomials of the 2nd degree of n variables. Then we can consider the following construction. Define new variables in a feature space:

$$\begin{aligned} z_1 = x_1, \ \ldots, \ z_n = x_n, \\ z_{n+1} = x_1^2, \ \ldots, \ z_{2n} = x_n^2, \\ z_{2n+1} = x_1 x_2, \ \ldots, \ z_N = x_n x_{n-1}. \end{aligned}$$

There are $N = 2n + \frac{n(n-1)}{2}$ of such variables. Thus, we have constructed the following non-linear mapping:

$$\bar{x} = (x_1, \ \ldots, \ x_n) \to \bar{\phi}(\bar{x}) = \bar{z} = (z_1, \ \ldots, \ z_N)$$

of the space $\mathcal{R}^n$ to the space $\mathcal{R}^N$.

A separating hyperplane in the feature space $Z = \mathcal{R}^N$:

$$(\bar{w} \cdot \bar{z}) + b = 0,$$

has a preimage in the initial space $\mathcal{R}^n$ that is a second-order hypersurface:

$$(\bar{w} \cdot \bar{\phi}(\bar{x})) + b = \sum_{i=1}^{N} w_i z_i + b =$$

$$= \sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{2n} w_i x_i^2 + \sum_{i=2n+1}^{N} w_i x_{j_i} x_{k_i} + b = 0,$$

where (j_i, k_i) is the ith pair of positive integer numbers in some one-one enumeration of all pairs of positive integers less than or equal to n.

Now consider the general case. Let a mapping (2.18)

$$\bar{\phi}(\bar{x}) = (\phi_1(\bar{x}),\ \ldots,\ \phi_N(\bar{x}))$$

of the space $\mathcal{R}^n$ to a feature space

$$\mathcal{R}^N = \{\bar{z} = (z_1,\ \ldots,\ z_N) : z_i \in \mathcal{R}, i = 1,\ \ldots,\ N\}$$

be given. In the coordinate form it can be written as $z_j = \phi_j(\bar{x})$, $j = 1,\ \ldots,\ N$.

The vectors $\bar{x}_1,\ \ldots,\ \bar{x}_l$ of the space $\mathcal{R}^n$ are mapped to the vectors $\bar{\phi}(\bar{x}_1),\ \ldots,\ \bar{\phi}(\bar{x}_l)$ of the feature space $\mathcal{R}^N$.

Using the method developed in Section 2.2, we construct the optimal hyperplane in the feature space $\mathcal{R}^N$:

$$\sum_{j=1}^{N} w_j z_j + b = 0, \tag{2.19}$$

separating the vectors $\bar{\phi}(\bar{x}_1),\ \ldots,\ \bar{\phi}(\bar{x}_l)$.

The preimage of this hyperplane in the space $\mathcal{R}^n$ is a non-linear hypersurface:

$$\sum_{j=1}^{N} w_j \phi_j(\bar{x}) + b = 0. \tag{2.20}$$

Using the dual form of the optimal hyperplane in the feature space we can represent its weight vector as a linear combination of support vectors from the set $\{\bar{\phi}(\bar{x}_i) : \alpha_i^0 > 0\}$:

$$\bar{w} = \sum_{i=1}^{l} \alpha_i^0 y_i \bar{\phi}(\bar{x}_i).$$

In the coordinate form, we have

$$w_j = \sum_{i=1}^{l} \alpha_i^0 y_i \phi_j(\bar{x}_i) \tag{2.21}$$

for $j = 1, \ldots, N$. The number of terms in this sum is independent of dimension of the feature space.

Substituting (2.21) in (2.20), we obtain a non-linear equation defining the preimage in $\mathcal{R}^n$ of the optimal separating hyperplane constructed in the feature space $\mathcal{R}^N$:

$$\begin{aligned}
&\sum_{j=1}^{N} w_j \phi_j(\bar{x}) + b = \\
&= \sum_{j=1}^{N} \left(\sum_{i=1}^{l} \alpha_i^0 y_i \phi_j(\bar{x}_i) \right) \phi_j(\bar{x}) + b = \\
&= \sum_{i=1}^{l} \alpha_i^0 y_i \sum_{j=1}^{N} \phi_j(\bar{x}) \phi_j(\bar{x}_i) + b = \\
&= \sum_{i=1}^{l} \alpha_i^0 y_i (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{x}_i)) + b = \\
&= \sum_{i=1}^{l} \alpha_i^0 y_i K(\bar{x}, \bar{x}_i) + b = 0,
\end{aligned} \tag{2.22}$$

where

$$K(\bar{x}_i, \bar{x}) = (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x})). \tag{2.23}$$

Thus, all statements about linear SVM (optimal hyperplanes in the space $\mathcal{R}^n$) hold also for non-linear SVM in the same space if we replace the dot product $(\bar{x}_i \cdot \bar{x})$ in the dual representation (2.16) :

$$\sum_{s=1}^{k} \alpha_i^0 y_i (\bar{x}_i \cdot \bar{x}) + b = 0$$

on the function $K(\bar{x}_i, \bar{x})$ defined by (2.23). We call it *the kernel*.

Note that the process of computation of the non-linear function

$$f(\bar{x}) = \sum_{i=1}^{l} \alpha_i^0 y_i K(\bar{x}_i, \bar{x}) + b, \tag{2.24}$$

corresponding to the hyperplane (2.22) requires only l operations and does not depend on the dimension N of the feature space. It is also clear, that to define a non-linear classifier in the space $\mathcal{R}^n$ we need not know the mapping $\bar{x} \to \bar{\phi}(\bar{x})$, and it is sufficient to know only the kernel $K(\bar{x}_i, \bar{x})$.

To solve the primal problem, we construct a hyperplane in $\mathcal{R}^N$ separating the images:

$$\bar{\phi}(\bar{x}_1), \ \ldots, \ \bar{\phi}(\bar{x}_l)$$

of vectors $\bar{x}_1, \ \ldots, \ \bar{x}_l$. To define the optimal hyperplane in $\mathcal{R}^N$, we have to solve the dual problem – to maximize the function $W(\alpha)$ defined by (2.13).

Using the kernel, we obtain

$$\begin{aligned} W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x}_j)) = \\ = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\bar{x}_i, \bar{x}_j) \end{aligned} \tag{2.25}$$

Thus, to find the optimal hypersurface (2.24) separating a sample $((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$ we have to maximize the non-linear function (2.25) under constraints (2.12) and $\alpha_i \geqslant 0$, $i = 1, \ \ldots, \ l$.

In this case, we do not require the knowledge of N-dimensional vectors $\bar{\phi}(\bar{x}_1),\ \ldots,\ \bar{\phi}(\bar{x}_l)$. It is enough to know their pairwise inner products defined by the kernel $K(\bar{x}_i, \bar{x}_j)$.

In practice, we choose a kernel for which the corresponding hypersurface best separates the training sample.

2.5. Kernels

In this section we consider the properties of kernels in more detail. Let X be an arbitrary set. In general, by a kernel we mean any function $K(x, y)$ mapping the set $X \times X$ into the set of all real numbers $\mathcal{R}$, which can be represented as a dot product:

$$K(x, y) = (\phi(x) \cdot \phi(y)), \tag{2.26}$$

where $\phi : X \to \mathcal{F}$ is a mapping from the set X into some feature space $\mathcal{F}$ endowed with the dot product.

Let us consider some examples of kernels, which are used in practical applications.

Example. $K(\bar{x}, \bar{y}) = (\bar{x} \cdot \bar{y})^d$ or $K(\bar{x}, \bar{y}) = ((\bar{x} \cdot \bar{y}) + c)^d$ – polynomial kernels.

Consider a mapping from $\mathcal{R}^n$ to $\mathcal{R}^{\frac{n(n+1)}{2}}$:

$$\bar{\phi}(\bar{x}) = \bar{\phi}(x_1, \ldots, x_n) =$$
$$= (1, x_1^2,\ \ldots,\ x_n^2, \sqrt{2}x_1x_2,\ \ldots,\ \sqrt{2}x_1x_2,\ \ldots,\ \sqrt{2}x_{n-1}x_n).$$

We have

$$K(\bar{x}, \bar{y}) = (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{y})) = 1 + \sum_{i=1}^{n} x_i^2 y_i^2 + \sum_{i,j=1, i<j}^{n} 2x_i x_j y_i y_j = (1 + \bar{x} \cdot \bar{y})^2.$$

Therefore, we have obtained the polynomial kernel of the second order: $K(\bar{x}, \bar{y}) = (1 + (\bar{x} \cdot \bar{y}))^2$.

The classifier (2.24) defined by the optimal separating hyperplane in the feature space $\mathcal{R}^{\frac{n(n+1)}{2}}$ has the form:

$$f(\bar{x}) = \sum_{i=1}^{l} \alpha_i^0 y_i (\bar{x}_i \cdot \bar{x})^2 + b. \tag{2.27}$$

Another set of kernels is defined by functions that have the form $K(\bar{x}, \bar{y}) = K(\bar{x} - \bar{y})$. Any such a function is invariant with respect to the addition to $\bar{x}$ and $\bar{y}$ of the same vector.

In case of dimension one, let a function $K(x)$ is defined on $[0, 2\pi]$. In this case, it can be extended to a periodic function on the real line and expanded in a uniformly convergent Fourier series:

$$K(x) = \sum_{n=0}^{\infty} a_n \cos(nx).$$

Then

$$K(x, y) = K(x - y) =$$
$$= a_0 + \sum_{n=0}^{\infty} a_n \sin(nx) \sin(ny) + \sum_{n=0}^{\infty} a_n \cos(nx) \cos(ny).$$

This kernel is defined by the mapping:

$$x \to (1, \sin x, \cos x, \sin 2x, \cos 2x, \ \ldots, \ \sin(nx), \cos(nx), \ \ldots)$$

of the initial space to a feature space.

In regression problems is widely used Gaussian kernel:

$$K(\bar{x}, \bar{y}) = \exp(-\|\bar{x} - \bar{y}\|^2 / \sigma^2).$$

Gaussian kernel can be obtained by transformation of the exponential kernel

$$K(\bar{u}, \bar{v}) = \exp((\bar{u} \cdot \bar{v}) / \sigma^2),$$

where $\sigma > 0$ is a parameter. This can be done as follows.

Exponential kernel can be expanded in Taylor series as the infinite sum of polynomial kernels:

$$\exp((\bar{u} \cdot \bar{v})) = \sum_{k=0}^{\infty} \frac{(\bar{u} \cdot \bar{v})^k}{k!}.$$

Exponential kernel is transformed into a Gaussian kernel as follows:

$$\frac{K(\bar{u},\bar{v})}{\sqrt{K(\bar{u},\bar{u})K(\bar{v},\bar{v})}} =$$
$$= \frac{\exp((\bar{u}\cdot\bar{v})/\sigma^2)}{\sqrt{\exp((\bar{u}\cdot\bar{u})/\sigma^2)\exp((\bar{v}\cdot\bar{v})/\sigma^2)}} =$$
$$= \exp(-\|\bar{u}-\bar{v}\|^2/2\sigma^2).$$

In the problem of text recognition, kernels defined on discrete sets are used. Here is an example of such a kernel and the corresponding feature space.

Let Ξ be a finite alphabet. A word s is a finite sequence of letters $s = s_1\ s_2\ \ldots\ s_n$; Ξ^* be a set of all words in the alphabet Ξ including empty set. Also, $|s| = n$ be the length of a word $s \in \Xi^*$, ie, the number of its letters; the length of empty set is 0.

Let Ξ^n be a set of all words of length n. By definition $\Xi^* = \cup_{n=0}^{\infty}\Xi^n$.

Let st be a concatenation of word s and t, $s[i:j] = s_i\ s_{i+1}\ \ldots\ s_j$.

A word u is a subword (subsequence) of a word s if a sequence of indices $\bar{i} = (i_1,\ \ldots,\ i_{|u|})$ exists such that $1 \leqslant i_1 < \cdots < i_{|u|} \leqslant |s|$ and $u_j = s_{i_j}$ for all $j = 1,\ \ldots,\ |u|$; this is also denote $u = s[\bar{i}]$. By the length of a subsequence u of a sequence s we mean a number $l(\bar{i}) = i_{|u|} - i_1 + 1$.

We assume that all words are linearly ordered: words of shorter length precedes to the words of greater length, and all the words of the same length are ordered lexicographically.

For any n, define a feature space $F_n = \mathcal{R}^{\Xi^n}$ and a mapping

$$\bar{\phi}^n(s) = (\phi_u^n(s) : u \in \Xi^n),$$

where

$$\phi_u^n(s) = \sum_{\bar{i}:u=s[\bar{i}]} \lambda^{l(\bar{i})}$$

for $0 < \lambda \leqslant 1$.

The corresponding kernel is defined as a dot product:

$$K_n(s,t) = \sum_{u \in \Xi^n} (\phi_u^n(s) \cdot \phi_u^n(t)) =$$
$$= \sum_{u \in \Xi^n} \sum_{\bar{i}: u = s[\bar{i}]} \lambda^{l(\bar{i})} \sum_{\bar{j}: u = s[\bar{j}]} \lambda^{l(\bar{j})} =$$
$$= \sum_{u \in \Xi^n} \sum_{\bar{i}: u = s[\bar{i}]} \sum_{\bar{j}: u = s[\bar{j}]} \lambda^{l(\bar{i}) + l(\bar{j})}.$$

This definition is computational inefficient: the direct computation of the kernel of $K_n(s,t)$ requires a huge number of computational operations. There is a recursive scheme for computation $K_n(s,t)$ in the polynomial time (see [10]).

For more details of the kernel theory see [30].

2.5.1. Positive semidefinite kernels

.

In this section we shall study the kernels of special type – positive semidefinite kernels. For any such kernel, a canonical Hilbert feature space can be constructed.

First consider the example from Section 2.4. Let a function $K(\bar{x}, \bar{y}) = (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{x}))$ be defined by some mapping $\bar{\phi}$ from the Euclidian space $\mathcal{R}^n$ to the Euclidian feature space $\mathcal{R}^N$.

By definition the function $K(\bar{x}, \bar{y})$ is symmetric: $K(\bar{x}, \bar{y}) = K(\bar{y}, \bar{x})$ for all $\bar{x}$ $\bar{y}$. Besides, another important property is valid: for any sequence $\bar{x}_1, \ldots, \bar{x}_n$ of vectors and for any sequence of real numbers $\alpha_1, \ldots, \alpha_n$:

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K(\bar{x}_i, \bar{x}_j) = \sum_{i,j=1}^{n} \alpha_i \alpha_j (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x}_j)) =$$
$$= \left(\sum_{i=1}^{n} \alpha_i \bar{\phi}(\bar{x}_i) \cdot \sum_{i=1}^{n} \alpha_i \bar{\phi}(\bar{x}_i) \right) = \left\| \sum_{i=1}^{n} \alpha_i \bar{\phi}(\bar{x}_i) \right\|^2 \geqslant 0. \qquad (2.28)$$

We state the property (2.28) as a definition. Let X be a set. A function $K : X \times X \to \mathcal{R}$ is called *positive semidefinite* if for any

collection $x_1, \ldots, x_n$ of vectors and for any sequence $\alpha_1, \ldots, \alpha_n$ of real numbers the following inequality is valid:

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \geqslant 0.$$

By (2.28) the function $K(\bar{x}, \bar{y}) = (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{x}))$ is positive definite.

A matrix $(K(x_i, x_j)_{i,j=1}^n$ is called Gram matrix of the kernel K.

Reproducing kernel Hilbert space

Let X be a set. For any symmetric positive semidefinite function $K(x, y)$ on $X \times X$ define a canonical functional Hilbert space $\mathcal{F}$. Define a mapping $\phi : X \to \mathcal{R}^X$ from the set X to the set of all functions from X to $\mathcal{R}$:

$$\phi(x) = K(\cdot, x) = K_x.$$

By definition $\phi(x)$ is a function, for which $K_x(y) = K(x, y)$ for all y. Consider a function space generated by all linear combinations:

$$f = \sum_{i=1}^{n} \alpha_i K_{x_i}, \tag{2.29}$$

for all n, $\alpha_i \in \mathcal{R}$, and $x_i \in X$.

Operations of sum and multiplication by a constant are defined in the standard way. The dot product of two functions $f = \sum\limits_{i=1}^{n} \alpha_i K_{x_i}$ and $g = \sum\limits_{j=1}^{n'} \beta_j K_{x'_j}$ is defined as

$$(f \cdot g) = \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \beta_j K(x_i, x'_j). \tag{2.30}$$

It is easy to verify that the expression (2.30) can be written as $(f \cdot g) = \sum\limits_{j=1}^{n'} \beta_j f(x'_j)$ or $(f \cdot g) = \sum\limits_{i=1}^{n} \alpha_i g(x_i)$.

It follows that the expression (2.30) is uniquely defined and does not depend on the representation of the functions f and g in the

form of linear combinations. It follows also that the function $(f \cdot g)$ is bilinear in f and g. It is also symmetric: $(f \cdot g) = (g \cdot f)$ for all f and g and positive semidefinite.

First, note that

$$(f \cdot f) = \sum_{i=1}^{n} \sum_{i,j=1}^{n'} \alpha_i \alpha_j K(x_i, x_j) \geqslant 0.$$

Taking into account this property, we obtain that for any set of functions $f_1 \ , \ldots , \ f_n$ and for any set $\alpha_1 \ , \ldots , \ \alpha_n \in \mathcal{R}$ of real numbers the condition

$$\sum_{i,j=1}^{n} \sum_{j=1}^{n'} \alpha_i \alpha_j (f_i \cdot f_j) = \left(\sum_{i=1}^{n} \alpha_i f_i \cdot \sum_{j=1}^{n} \alpha_j f_j \right) \geqslant 0$$

is valid. This means that $(f \cdot g)$ is positive semidefinite function.

By (2.30) the identity $(K_x \cdot f) = f(x)$ holds. In particular, $(K_x \cdot K_y) = K(x, y)$ for all x and y. By this reason, any positive semidefinite symmetric function $K(x, y)$ is called *reproducing kernel*.

From this and by (2.29), see also (2.89) in a problem of Section 2.12, we have

$$|f(x)|^2 = |(f \cdot K_x)|^2 \leqslant K(x, x)(f \cdot f).$$

In particular, $(f \cdot f) = 0$ implies that $f(x) = 0$ for all x.

The function $\|f\| = \sqrt{(f \cdot f)}$ is a norm, since it is defined by the dot product.

Consider the completion of the set of all linear combinations of (2.29) with respect to this norm to a complete metric space $\mathcal{F}$.

The obtained canonical Hilbert space is called *Reproducing Kernel Hilbert Space – RKHS*.

An alternative variant of RKHS definition is as follows. RKHS is a Hilbert space $\mathcal{F}$ of functions on X, which has the following property: the functional $f \to f(x)$ is a continuous linear functional. By Riesz–Fisher theorem for each $x \in X$ there exists $K_x \in \mathcal{F}$ such that $f(x) = (K_x \cdot f)$. The reproducing kernel is defined $K(x, y) = (K_x \cdot K_y)$.

The Gaussian kernel $K(\bar{x}, \bar{y}) = \exp(-\|\bar{x} - \bar{y}\|^2/\sigma^2)$ is also positive semidefinite, and so, defines a canonical Hilbert space and a corresponding mapping.

We note without proof that although the kernel defined by some mapping ϕ in a feature space, using the equation (2.26), is symmetric and positive semidefinite the canonical Hilbert space $\mathcal{F}$ and the corresponding mapping may be different from the original feature space and mapping ϕ. On the other hand, each symmetric positive semidefinite kernel defines a unique RKHS (see [2]).

Representer theorem

Theorem of the representative (Representer theorem) shows that solutions of a wide class of optimization problems can be expressed as linear combinations of values of kernels in the training data points. This theorem was proved by Kimmeldorf and Wahba [20]. See also [30].

Theorem 2.3. *Let X be a set and $S = ((x_1, y_1) \ , \ldots , \ (x_l, y_l))$ be a training sample, where $(x_i, y_i) \in X \times \mathcal{R}$. Let also, $K(x, x')$ be a positive semidefinite kernel on $X \times X$ and $\mathcal{F}$ be the corresponding canonical RKHS with a norm $\| \cdot \|$.*

A loss function $c : (X^2 \times \mathcal{R})^l \to \mathcal{R} \cup \{\infty\}$ and strictly monotonic function Ω defined on $\mathcal{R}$ are given.

Then any function $f \in \mathcal{F}$ minimizing the regularized risk functional

$$c((x_1, y_1, f(x_1)) \ , \ldots , \ (x_l, y_l, f(x_l))) + \Omega(\|f\|) \tag{2.31}$$

can be represented as

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) \tag{2.32}$$

for some real numbers $\alpha_1 \ , \ldots , \ \alpha_l$.

An example of such functional for regression problem in a feature space $f \in \mathcal{F}$ is:

$$c((x_1, y_1, f(x_1)) \ , \ldots , \ (x_l, y_l, f(x_l))) = \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2 + \lambda \|f\|^2,$$

where $\lambda > 0$.

Proof. Recall that $K_{x_i} = K(x_i, \cdot)$ is a function defined by the kernel $K(x, y)$. Any function $f \in \mathcal{F}$ can be represented as $f(x) = (f \cdot K_x)$ for all x.

Consider a decomposition of the linear space $\mathcal{F}$ into a direct sum of a finite space generated by all linear combinations of the functions K_{x_i}, $i = 1, \ldots, l$, and its orthogonal complement. Then any function $f \in \mathcal{F}$ is represented as:

$$f = \sum_{i=1}^{l} \alpha_i K_{x_i} + f_*,$$

where $(f_* \cdot K_{x_i}) = 0$ for all $i = 1, \ldots, l$.

Compute the values $f(x_j)$ for all $j = 1, \ldots, l$:

$$f(x_j) = (f \cdot K_{x_j}) =$$
$$= \left(\left(\sum_{i=1}^{l} \alpha_i K_{x_i} + f_* \right) \cdot K_{x_j} \right) =$$
$$= \sum_{i=1}^{l} \alpha_i (K_{x_i} \cdot K_{x_j}).$$

It is important that the value of $f(x_j)$ does not depend on the element f_* from the orthogonal complement. Thus, the value of the main part $c((x_1, y_1, f(x_1)), \ldots, (x_l, y_l, f(x_l)))$ of the regularized functional (2.31) does not depend on f_*.

Since f_* is orthogonal to the element $\sum_{i=1}^{l} \alpha_i K_{x_i}$ and the function Ω is strictly monotone, the following inequality holds:

$$\Omega(\|f\|) = \Omega \left(\| \sum_{i=1}^{l} \alpha_i K_{x_i} + f_* \| \right) =$$
$$= \Omega \left(\sqrt{\| \sum_{i=1}^{l} \alpha_i K_{x_i} \|^2 + \|f_*\|^2} \right) \geqslant$$
$$\geqslant \Omega \left(\| \sum_{i=1}^{l} \alpha_i (K_{x_i} \| \right),$$

with equality if and only if $f_* = 0$. Therefore, at the minimum point of the functional (2.31) it must be $f_* = 0$.

Hence, the solution of the problem of minimizing the functional (2.31) must have the form (2.32):

$$f = \sum_{i=1}^{l} \alpha_i K_{x_i}.$$

Theorem is proved. $\triangle$

Theorem 2.3 shows that to obtain a solution of the problem (2.31) of functional minimization in any RKHS (which may be infinite dimensional), it is sufficient to solve the minimization problem in a finite dimensional space $\mathcal{R}^n$.

An example of a risk functional corresponding to the optimization problem for SVM:

$$\begin{aligned} c((x_1, y_1, f(x_1)) \ , \ldots , \ (x_l, y_l, f(x_l))) = \\ = \frac{1}{\lambda} \sum_{i=1}^{l} \max\{0, 1 - y_i f(x_i)\} + \|f\|^2, \end{aligned}$$

where $x_i \in \mathcal{R}^n$ and $y_i \in \{-1, +1\}$ for $i = 1 \ , \ldots , \ l$.

The corresponding feature space $\mathcal{F}$ is generated by the kernel $K(x, x')$. A function $f \in \mathcal{F}$ minimizing the risk functional (2.31) has the form

$$f = \sum_{i=1}^{l} \alpha_i K_{x_i}.$$

2.6. Inseparable training sample

First, we obtain an upper bound for the generalization error in the case when a sample is not completely separated by a classification function. This estimate serves as a basis for setting the corresponding optimization problem of constructing an optimal classifier.

2.6.1. Margin slack variables

Now consider the problem of classification of an inseparable sample. Problems of this type are typical for practical applications.

Let a class $\mathcal{F}$ of functions of type $\mathcal{X} \to \mathcal{R}$ be given. Any such a function $f \in \mathcal{F}$ defines a classifier:

$$h(x) = \begin{cases} 1 \text{ if } f(x) \geqslant 0, \\ -1 \text{ otherwise.} \end{cases}$$

Let a sample $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$ be given and $\gamma_i = y_i f(x_i)$ be the margin of an example $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ with respect to a function $f \in \mathcal{F}$.

Note that $\gamma_i > 0$ means that the classification by f is correct.

The margin distribution of f with respect to the training set $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$ is defined by the vector $M_S(f) = (\gamma_1, \ \ldots, \ \gamma_l)$. We refer to the minimum of the margin distribution as the margin

$$m_S(f) = \min_{i=1,\ldots,l} \gamma_i$$

of f with respect to the sample S. Eviently, $m_S(f) > 0$ if and only if when f strongly separates S.

The margin of a training set S with respect to the class $\mathcal{F}$ is the maximal margin over all $f \in \mathcal{F}$.

We define *the margin slack variable* of an example $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ with respect to a function $f \in \mathcal{F}$ and target margin $\gamma > 0$ to be the quantity

$$\xi_i = \max\{0, \gamma - y_i f(\bar{x}_i)\}.$$

This is the amount by which the function f fails to achieve margin γ for the example (x_i, y_i).

A vector $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ is called *margin slack vector* of a training set $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$. By definition $y_i f(x_i) + \xi_i \geqslant \gamma$ for all i.

If $\xi_i > \gamma$ then $y_i f(\bar{x}_i) < 0$, ie, classification of the example (x_i, y_i) by f is incorrect. In this case, the quantity of ξ_i reflects the value of remoteness of the example (x_i, y_i) from the separating hyperplane – it is greater, the greater the error of classification.

It hold $\xi_i = 0$ if and only if $y_i f(x_i) \geqslant \gamma$; in this case classification is correct and even with some margin.

The case $0 < \xi_i \leqslant \gamma$ is intermediate; in this case the classification $0 < y_i f(x_i) \leqslant \gamma$ is correct, but with a very low threshold. For example, it may be because of presence of noise in the original data.

In general, the norm of the error vector $\bar{\xi}$ reflects the number and the value of classification errors, and also the role of noise in the training set. In what follows, a value $\|\bar{\xi}\|$ will be part of the upper bounds of generalization error.

If the norm of vector $\bar{\xi}$ is positive the training sample is inseparable by the classifier $f(\bar{x})$ with a threshold of $\gamma > 0$. Theorem 1.9 is not directly applicable in this case.

However, in the case of a linear classifier on $\mathcal{R}^n$ we can replace this problem by equivalent one in a space of higher dimension, where a modified training set is separable. The corresponding result of Shawe-Taylor and Cristianini [27] is presented in the following theorem.

Theorem 2.4. *Let $\gamma > 0$ and $\mathcal{L}$ be a class of all linear homogeneous functions on $\mathcal{R}^n$. $L(\bar{x}) = (\bar{w} \cdot \bar{x})$ with a unit weight vector $\|\bar{w}\| = 1$. Let also, P be a continuous probability distribution on $\mathcal{X} \times \{-1, 1\}$ with a support into a ball of radius R centered at the origin.* [1]

Then for any $\delta > 0$, with probability $1-\delta$, any classifier $f \in \mathcal{L}$ with the margin slack variable ξ satifying $\|\xi\| \leqslant \|\xi_0\|$ has a generalization error

$$\operatorname{err}_P(f) \leqslant \frac{c}{l}\left(\frac{R^2 + \|\bar{\xi}_0\|^2}{\gamma^2}\log^2 l + \log\frac{2}{\delta}\right), \tag{2.33}$$

where $c > 0$ is a constant and $\bar{\xi}_0$ is the upper bound of the margin slack vectors with respect to functions from $\mathcal{L}$ and a target margin $\gamma > 0$.

Proof. Let $f(\bar{x}) = (\bar{w} \cdot \bar{x})$ be a classifier such that $\|\bar{w}\| = 1$. By definition of the margin slack variable $\bar{\xi} = (\xi_1, \ldots, \xi_l)$ defined for

[1] Here we mean a ball in $\mathcal{R}^n$ containing elements $\bar{x}_1, \ldots, \bar{x}_l$ of the training sample. A probability distribution is continuous if probability of any pair $(\bar{x}, y)$ is 0.

a sample $\bar{x}_1, \ldots, \bar{x}_l$ with respect to a function L and target margin $\gamma > 0$

$$y_i f(\bar{x}_i) + \xi_i \geqslant \gamma \tag{2.34}$$

for $i = 1, \ldots, l$.

Let $\nu > 0$ be a parameter whose value we will optimize later. Replace the training vectors $\bar{x}_1, \ldots, \bar{x}_l$ of dimension n on auxiliary vectors $\bar{x}'_1, \ldots, \bar{x}'_l$ of dimension $n + l$ defined as

$$\bar{x}'_i = (x_{i,1}, \ldots, x_{i,n}, 0, \ldots, \nu, \ldots, 0)$$

for $i = 1, \ldots, l$, where $(n + i)$th coordinate of the vector $\bar{x}'_i$ is ν and other extra coordinates are 0. The resulting sample is denoted $S' = ((\bar{x}'_1, y_1) \ldots, (\bar{x}'_l, y_l))$. We extend all other vectors by l zeros.

The hyperplane $f(\bar{x}) = (\bar{w} \cdot \bar{x})$ is replaced by the hyperplane

$$f'(\bar{x}') = (\bar{w}' \cdot \bar{x}'), \text{ where}$$
$$\bar{w}' = (w_1, \ldots, w_n, \frac{1}{\nu} y_1 \xi_1, \ldots, \frac{1}{\nu} y_l \xi_l), \tag{2.35}$$

and $\bar{x}'$ is a vector of dimension $n + l$.

By (2.34) a new sample S' is separated by the new classifier (2.35) with the margin γ:

$$y_i(\bar{w}' \cdot \bar{x}'_i) = y_i(\bar{w} \cdot \bar{x}_i) + (y_i)^2 \xi_i \geqslant \gamma \tag{2.36}$$

for $i = 1, \ldots, l$.

The classifiers f f' work in the same way outside S and S': $yf'(\bar{x}') = yf(\bar{x})$. Then error probabilities of these classifiers are the same.

In order to apply Theorem 1.9 from Section 1.3 to the new sample and new classifier, it is necessary to normalize the weight vector of the hyperplane (2.35). We have

$$\|\bar{w}'\|^2 = \|\bar{w}\|^2 + \frac{1}{\nu^2}\|\bar{\xi}\|^2 = 1 + \frac{1}{\nu^2}\|\bar{\xi}\|^2.$$

Besides, by definition all vectors $\bar{x}'_i$ belong to the ball of radius R', where $R'^2 = R^2 + \nu^2$.

After normalization, condition (2.36) reduces to the condition

$$y_i\left(\frac{\bar{w}'}{\|\bar{w}'\|}\cdot\bar{x}_i'\right)\geqslant\gamma'=\frac{\gamma}{\|\bar{w}'\|}.$$

for $i=1,\ \ldots,\ l$.

From this follows that the main factor of the bound of Corollary 1.4 has the form

$$\frac{R'^2}{\gamma'^2}=\frac{(R^2+\nu^2)(1+\frac{1}{\nu^2}\|\xi\|^2)}{\gamma^2}.$$

Transform it to the expression

$$(R^2+\nu^2)(1+\frac{1}{\nu^2}\|\xi\|^2)=R^2+\|\xi\|^2+\nu^2+\frac{1}{\nu^2}R^2\|\xi\|^2.$$

A minimum of this expression is attained at $\nu^2=R\|\xi\|$, and the expression takes the form

$$R^2+2R\|\xi\|+\|\xi\|^2=(R+\|\xi\|)^2\leqslant 2(R^2+\|\xi\|^2).$$

Radius if a new ball is defined $R'=\sqrt{2(R^2+\|\xi_0\|^2)}$, where $\|\xi_0\|$ – is the upper bound of norms of margin slack variables.

Applying Theorem 1.9 from Section 1.3, we obtain a bound (2.33). Theorem is proved. $\triangle$

2.6.2. Soft margin optimization

Quadratic norm optimization

In case where a training sample $S=((\bar{x}_1,y_1),\ \ldots,\ (\bar{x}_l,y_l))$ is inseparable the optimization problem with a margin slack vector ξ_i, $i=1,\ \ldots,\ l$ is considered.

In the soft setting we allow the margin constrains to be violated.

Search for vectors $\bar{w}$, $\bar{\xi}$ and a number b such that

$$(\bar{w}\cdot\bar{w})+C\sum_{i=1}^{l}\xi_i^2\to\min, \tag{2.37}$$

$$y_i((\bar{w}\cdot\bar{x}_i)+b)\geqslant 1-\xi_i, \tag{2.38}$$

$$\xi_i\geqslant 0 \tag{2.39}$$

for $i = 1, \ \ldots, \ l$. A constant C defines a balance between two parts of the functional.

In practice the parameter C is varied through a wide range of values and the optimal performance assessed using a separate validation set or the technics of cross-validation for verifying the performance using only the training sample.

Note that the condition $\xi_i \geqslant 0$ can be dropped, since the optimal solution $\bar{w}$, ξ, b, where $\xi_i < 0$ for some i, is also optimal for $\xi_i = 0$.

The Lagrangian of the primal problem (2.37) – (2.39) is

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) + \frac{C}{2}\sum_{i=1}^{l} \xi_i^2 - \\ - \sum_{i=1}^{l} \alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i), \tag{2.40}$$

where $\alpha_i \geqslant 0$ are the Lagrange multipliers.

The corresponding dual problem is solved by differentiating by $\bar{w}$, $\bar{\xi}$ and b:

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} y_i \alpha_i \bar{x}_i = \bar{0}, \tag{2.41}$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial \bar{\xi}} = C\bar{\xi} - \bar{\alpha} = \bar{0},$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0,$$

and substituting the relations obtained into the primal (2.40) to obtain the following adaptation of the dual objective functions:

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j) + \\ + \frac{1}{2C}(\bar{\alpha} \cdot \bar{\alpha}) - \frac{1}{C}(\bar{\alpha} \cdot \bar{\alpha}) = \\ = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j) - \frac{1}{2C}(\bar{\alpha} \cdot \bar{\alpha}). \tag{2.42}$$

After that, we maximize over $\bar{\alpha}$ the quantity

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j ((\bar{x}_i \cdot \bar{x}_j) + \frac{1}{C} \delta_{ij}) \tag{2.43}$$

under constraints $\alpha_i \geqslant 0$, $i = 1, 2, \ldots l$, where $\delta_{ij} = 1 \;\; i = j$ and $\delta_{ij} = 0$ for $i \neq j$. The corresponding Karush–Kuhn–Tucker conditions are:

$$\alpha_i (y_i ((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i) = 0$$

for $i = 1, \;\ldots,\; l$.

By (2.41) the weight vector is a linear combination of support vectors:

$$\bar{w} = \sum_{i=1}^{l} y_i \alpha_i \bar{x}_i.$$

From the Karush–Kuhn–Tucker conditions follows that $\alpha_i = 0$ and $\xi_i = 0$ if $y_i ((\bar{w} \cdot \bar{x}_i) + b) > 1$.

These vectors are correctly classified and lie on the outside of the boundary hyperplanes. Support vectors are those vectors $\bar{x}_i$ for which $y_i ((\bar{w} \cdot \bar{x}_i) + b) \leqslant 1$, with $\alpha_i \geqslant 0$ and $\xi_i \geqslant 0$. These are those vectors that lie on the boundary hyperplanes or incorrectly classified by them, in this case $y_i ((\bar{w} \cdot \bar{x}_i) + b) < 1 \;\; \xi_i > 0$.

Let us formulate the optimization problem for a feature space defined by some kernel $K(\bar{x}_i, \bar{x}_j)$.

Theorem 2.5. *Assume that a feature space defined by a kernel $K(\bar{x}_i, \bar{x}_j)$ and a training sample $S = ((\bar{x}_1, y_1), \;\ldots,\; (\bar{x}_l, y_l))$ be given. Let the parameter vector $\bar{\alpha}^*$ be a solution of the optimization problem:*

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i -$$

$$-\frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (K(\bar{x}_i, \bar{x}_j) + \frac{1}{C} \delta_{ij}) \to \max \tag{2.44}$$

$$\textit{subject to } \sum_{i=1}^{l} y_i \alpha_i = 0,$$

$$\alpha_i \geqslant 0 \;\; i = 1, \;\ldots,\; l. \tag{2.45}$$

Then the corresponding separating hypersurface is given by

$$f(\bar{x}) = \sum_{i=1}^{l} y_i \alpha_i^* K(\bar{x}_i, \bar{x}) + b^*,$$

where b^* *is chosen so that* $y_i f(\bar{x}_i) = 1 - \alpha_i^*/C$ *for any* i *such that* $\alpha_i^* \neq 0$.

The classifier $\mathrm{sign}(f(\bar{x}))$ *separates the training set as well as the corresponding hyperplane in the feature space implicitly defined by the kernel* $K(\bar{x}, \bar{z})$*, where the slack variables are defined relative to the margin*

$$\gamma = \left(\sum_{j \in sv} \alpha_j^* - \frac{1}{C} (\bar{\alpha}^* \cdot \bar{\alpha}^*) \right)^{-1/2}.$$

The value of b^* is chosen using the relation $\alpha_i = C\xi_i$ and by reference to the primal constraints defined by the Karush–Kuhn–Tucker conditions:

$$\alpha_i (y_i ((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i) = 0$$

for $i = 1, \ \ldots, \ l$.

The quantity $\rho(\bar{w}) = \frac{1}{|\bar{w}|}$ which determines the size of the margin $(\bar{w} \cdot \bar{x}_i) + b) = \pm 1$ is defined

$$\begin{aligned}
(\bar{w} \cdot \bar{w}) &= \sum_{i,j=1}^{l} y_i y_j \alpha_i^* \alpha_j^* K(\bar{x}_i, \bar{x}_j) = \\
&= \sum_{j \in sv} y_j \alpha_j^* \sum_{i \in sv} y_i \alpha_i^* K(\bar{x}_i, \bar{x}_j) = \\
&= \sum_{j \in sv} \alpha_j^* (1 - \xi_j^* - y_j b^*) = \\
&= \sum_{j \in sv} \alpha_j^* - \sum_{j \in sv} \alpha_j^* \xi_j^* = \\
&= \sum_{j \in sv} \alpha_j^* - \frac{1}{C} (\bar{\alpha}^* \cdot \bar{\alpha}^*).
\end{aligned}$$

The upper bound (2.33) of the generalization error does not depend on the dimension that allows us to apply this bound to the case of separation with a kernel $K(\bar{x}, \bar{z})$ generating a feature space of high dimension.

Now we apply Theorem 2.4. The bound (2.33) of the generalization error holds for classifiers defined by linear functions with unit weight vectors $\bar{w}$. In order to apply it for the problem (2.37) – (2.39), divide both sides of the inequality (2.38) on $\|\bar{w}\|$, where $\bar{w}$ is the optimal solution of the problem. We get the quantity $\xi_i/\|\bar{w}\|$ as ξ_i in (2.33). Define also $\gamma = 1/\|\bar{w}\|$. Then we obtain the inequality

$$\begin{gathered} \mathrm{err}_P(f) = P\{yf(\bar{x}) < 0\} \leqslant \\ \leqslant \frac{c}{l}((\|\bar{w}\|^2 R^2 + \|\bar{\xi}\|^2) \ln^2 l + \ln \frac{1}{\delta}). \end{gathered} \tag{2.46}$$

The inequality (2.46) shows that in order to minimize the upper bound of the generalization error we really need to minimize a value (2.37).

Linear norm optimization

It is also often an analogous optimization problem is considered in which, instead of the quadratic norm of a vector of margin slack variables $\bar{\xi}$, the linear norm is used. In this case, we have the following optimization problem.

Search for vectors $\bar{w}$, $\bar{\xi}$ and a number b such that

$$(\bar{w} \cdot \bar{w}) + C \sum_{i=1}^{l} \xi_i \to \min, \tag{2.47}$$

$$y_i((\bar{w} \cdot \bar{x}_i) + b_0) \geqslant 1 - \xi_i,$$

$$\xi_i \geqslant 0 \tag{2.48}$$

for $i = 1, \ \ldots, \ l$. The constant C determines the balance between two parts of the functional.

The corresponding Lagrangian is

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}, \bar{r}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) + C \sum_{i=1}^{l} \xi_i -$$

$$- \sum_{i=1}^{l} \alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b_0) - 1 + \xi_i) - \sum_{i=1}^{l} r_i \xi_i,$$

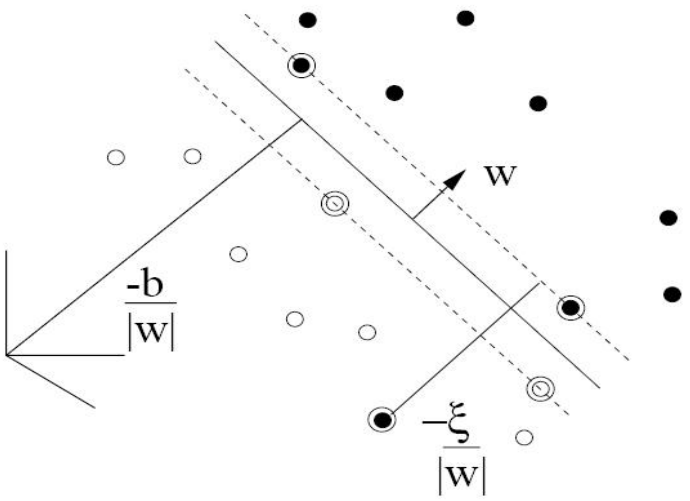


Fig. 1.2. Support vectors are located on the boundary hyperplanes or incorrectly classified by them

where $\alpha_i \geqslant 0$, $r_i \geqslant 0$ for $i = 1, \ \ldots, \ l$.

The correspondent dual problem is obtained by equating to zero the derivatives:

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha}, \bar{r})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} y_i \alpha_i \bar{x}_i = \bar{0},$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha}, \bar{r})}{\partial \xi_i} = C - \alpha_i - r_i = 0,$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha}, \bar{r})}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0.$$

Substituting the solution of these equations in the direct problem, we obtain the dual representation of the problem as a maximization

problem:

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}, \bar{r}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j),$$

which is almost identical with the functional for the case of quadratic norm. The only difference from the problem for the quadratic norm is that the condition $C - \alpha_i - r_i = 0$ together with the condition $r_i \geqslant 0$ forces the inequality $\alpha_i \leqslant C$. At the same time, $\xi_i > 0$ is only if $r_i = 0$. This implies that $\alpha_i = C$ for all such i. Thus, the Karush–Kuhn–Tucker conditions have the form:

$$\alpha_i(y_i((\bar{x}_i \cdot \bar{x}_j) + b) - 1 + \xi_i) = 0, \ i = 1, \ \ldots, \ l,$$
$$\xi_i(\alpha_i - C) = 0, \ i = 1, \ \ldots, \ l.$$

According to these conditions the margin slack variable ξ_i is nonzero only when $\alpha_i = C$.

Support vectors – are those vectors $\bar{x}_i$ for that $\alpha_i > 0$ (in this case $\alpha_i = C$). For them, $y_i((\bar{x}_i \cdot \bar{x}_j) + b) \leqslant 1$ and $\xi_i \geqslant 0$. It is easy to see that the distance from this vector to the corresponding separating hyperplane is equal to $-\frac{\xi_i}{\|\bar{w}\|}$ (see Fig. 1.2).

For an arbitrary kernel the following statement holds.

Theorem 2.6. *Assume that a training sample* $S = ((\bar{x}_1, y_1), \ldots, (\bar{x}_l, y_l))$ *and a feature space defined by a kernel* $K(\bar{x}_i, \bar{x}_j)$ *be given. Let also, a parameter vector* $\bar{\alpha}^*$ *is a solution of the optimization problem*

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(\bar{x}_i, \bar{x}_j) \to \max \qquad (2.49)$$

$$\text{subject to } \sum_{i=1}^{l} y_i \alpha_i = 0,$$
$$C \geqslant \alpha_i \geqslant 0 \ i = 1, \ \ldots, \ l. \qquad (2.50)$$

Then the corresponding separating hypersurface is given by

$$f(\bar{x}) = \sum_{i=1}^{l} y_i \alpha_i^* K(\bar{x}_i, \bar{x}) + b^*,$$

where b^* can be found from the condition $y_i f(\bar{x}_i) = 1$ for any i such that $C > \alpha_i^* > 0$.

Then the classifier $\operatorname{sign}(f(\bar{x}))$ *separates elements of a sample as well as the corresponding hyperplane obtained as the solution of the optimization problem (2.47) – (2.48) in the feature space defined by the kernel $K(\bar{x}, \bar{z})$, where the margin slack variables are defined for the margin*

$$\gamma = \left(\sum_{i,j \in sv} y_i y_j \alpha_i^* \alpha_j^* K(\bar{x}_i, \bar{x}_j) \right)^{-1/2}.$$

Thus, the optimization problem (2.47) – (2.48) is equivalent to the optimization problem (2.37) – (2.39) with an additional condition that $\alpha_i \leqslant C$. For this reason, these restrictions are called box constraints, since they require that all α_i be inside a square with side C located in the positive octant.

The parameter C controls the ratio between the accuracy of the regularization coefficient and value of α_i. In particular, the smaller the parameter C, the less the value of α_i, ie, the less influence of examples that are far from separating hyperplane.

Soft margin optimization as a linear programming problem

The previous problem can be formulated as a linear programming problem, in which, instead of the quadratic norm of the vector $\bar{w}$ the sum of the coefficients α_i is minimized. These coefficients characterize the degree of participation of examples in constructing the separating hyperplane.

The upper bound of the generalization error through the number of support vectors, given in Theorem 2.2, can serve as a justification of the applicability of the method.

In this case, we consider the problem of optimization:

$$\sum_{i=1}^{l} \alpha_i + C \sum_{i=1}^{l} \xi_i \to \min$$

$$\text{subject to } y_i \left(\sum_{j=1}^{l} \alpha_i (\bar{x}_i \cdot \bar{x}_j) + b \right) \geqslant 1 - \xi_i,$$

$$\alpha_i \geqslant 0,\ \xi_i \geqslant 0,$$

where $i = 1,\ \ldots,\ l$. The constant C determines a balance between two parts of the functional.

The advantage of this setting is that here the linear programming problem is solved instead of a quadratic programming problem.

2.7. Rademacher averages and generalization error

In this section, we present upper bounds of the generalization error for the classification functions defined by threshold functions from RKHS.[2]

Let $\mathcal{F}$ be an Hilbert space of functions defined on some set $\mathcal{X}$. We also assume that this space is RKHS, ie, it is generated by a reproducing kernel $K(x, y)$. Any function $f \in \mathcal{F}$ is represented as a scalar product $f(x) = (f \cdot \phi(x))$, where $\phi(x) = K(x, \cdot)$.

An example of such RKHS can be defined by the mapping $\phi : \mathcal{R}^n \to \mathcal{R}^N$. Let $\mathcal{F}$ be a space of functions $f(\bar{x}) = (\bar{w} \cdot \phi(\bar{x}))$, where $\bar{x} \in \mathcal{R}^n$, $\bar{w} \in \mathcal{R}^N$ and $(\bar{w} \cdot \bar{w}')$ is the dot product in $\mathcal{R}^N$. The norm of f is defined $\|f\| = \|\bar{w}\|$, and the scalar product of functions f and $g(\bar{x}) = (\bar{w}' \cdot \phi(\bar{x}))$ is defined $(f \cdot g) = (\bar{w} \cdot \bar{w}')$. The function $K(\bar{x}, \bar{y}) = (\phi(\bar{x}) \cdot \phi(\bar{y}))$ is the corresponding kernel.

Any function $f \in \mathcal{F}$ defines the classifier

$$h(\bar{x}) = \begin{cases} 1 \text{ if } f(x) \geqslant 0, \\ -1 \text{ otherwise.} \end{cases}$$

[2]The material in this section uses the results of Bartlett and Mendelson [3] and Shaw-Taylor and Cristianini [28].

Let $\mathcal{F}_1 = \{f \in \mathcal{F} : \|f\| \leqslant 1\}$. In the example above $\mathcal{F}_1$ is the class of functions $f(\bar{x}) = (\bar{w} \cdot \phi(\bar{x}))$ such that $\|\bar{w}\| \leqslant 1$.

Assume that a training set $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$ be given, where $x_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$.

Let $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$ be the Gram matrix defined by the values of the kernel on objects of the sample S; $tr(\mathbf{K}) = \sum_{i=1}^l K(x_i, x_i)$ be the trace of the matrix $\mathbf{K}$.

Now we estimate the empirical Rademacher average of the class $\mathcal{F}_1$ relative to the training set S.

Theorem 2.7. *The empirical Rademacher average of the class $\mathcal{F}_1$ relative to the training set $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$ satisfies the inequality:*

$$\tilde{\mathcal{R}}_l(\mathcal{F}_1) \leqslant \frac{1}{l}\sqrt{tr(\mathbf{K})}. \tag{2.51}$$

Proof. The following chain of equalities and inequalities is valid:

$$\begin{aligned}
\tilde{\mathcal{R}}_l(\mathcal{F}_1) = E_\sigma \left(\sup_{f \in \mathcal{F}_1} \left| \frac{1}{l} \sum_{i=1}^l \sigma_i f(x_i) \right| \right) = \\
= E_\sigma \left(\sup_{\|f\| \leqslant 1} \left| \left(f \cdot \frac{1}{l} \sum_{i=1}^l \sigma_i \phi(x_i) \right) \right| \right) \leqslant \\
\leqslant \frac{1}{l} E_\sigma \left(\left\| \sum_{i=1}^l \sigma_i \phi(\bar{x}_i) \right\| \right) = \\
= \frac{1}{l} E_\sigma \left(\left(\sum_{i=1}^l \sigma_i \phi(x_i) \cdot \sum_{i=1}^l \sigma_i \phi(x_i) \right)^{1/2} \right) \leqslant \\
\leqslant \frac{1}{l} \left(E_\sigma \left(\sum_{i,j=1}^l \sigma_i \sigma_j K(x_i, x_j) \right) \right)^{1/2} = \\
= \frac{1}{l} \left(\sum_{i=1}^l K(x_i, x_i) \right)^{1/2}.
\end{aligned}$$

Here in the transition from the 2nd line to the third the Cauchy–Schwarz inequality was used, in transition from the third row to

the 4th the definition of the norm vector was used. In the transition from the 4th row to the 5th Jensen inequality was used, in the transition from the 5th row to the 6th, we have used the independence of the random variables σ_i and equalities $E(\sigma_i^2) = 1$ and $E(\sigma_i\sigma_j) = E(\sigma_i)E(\sigma_j) = 0$ for $i \neq j$. Theorem is proved. $\triangle$

Recall that a number $\gamma_i = y_i f(x_i)$ is called the margin of an example $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ with respect to a function $f \in \mathcal{F}$. Note that if $\gamma_i > 0$ then the classification using f is correct.

Let a sample $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$ and a number $\gamma > 0$ be given. The margin slack variable for a function f and a margin bound γ is defined:

$$\xi_i = \max\{(0, \gamma - y_i f(x_i))\}. \tag{2.52}$$

Recall that if $\xi_i > \gamma$ then the classification of the example (x_i, y_i) is wrong.

Let $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ be the margin slack vector of a training set $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$.

Define an auxiliary function $f(x, y) = -yf(x)$ and the corresponding class of functions with domain $\mathcal{X} \times \{-1, 1\}$:

$$\mathcal{F}_2 = \{f(x, y) : f(x, y) = -yf(x), f \in \mathcal{F}_1\}.$$

Let

$$\chi(x) = \begin{cases} 1 \text{ if } x \geqslant 0, \\ 0 \text{ otherwise.} \end{cases}$$

Assume that examples (x_i, y_i) of the training set S are generated i.i.d. by some probability distribution P. It is easy to verify that

$$P\{(x, y) : y \neq \text{sign}(f(x))\} \leqslant E_P(\chi(-yf(x))).$$

Let $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$ be the Gram matrix defined by the kernel and the training set S.

The following theorem gives an upper bound for the generalization error of the classifier defined by the kernel K.

Theorem 2.8. *For any $\delta > 0$ and l, with probability $1 - \delta$, for any function $f \in \mathcal{F}_1$:*

$$P\{y \neq \operatorname{sign}(f(x))\} \leqslant \frac{1}{l\gamma} \sum_{i=1}^{l} \xi_i + \frac{2}{l\gamma} \sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \quad (2.53)$$

Note that the right side of (2.53) is a random variable, since the margin slack variables ξ_i are depend on $\bar{x}_i$.

Proof. Recall that $\gamma > 0$ is a margin bound. Define the auxilliary function $g : \mathcal{R} \to [0, 1]$:

$$g(r) = \begin{cases} 1 \text{ if } r > 0, \\ 1 + r/\gamma \text{ if } -\gamma \leqslant r \leqslant 0, \\ 0 \text{ otherwise.} \end{cases}$$

Since $g(r) \geqslant \chi(r)$ for all r, and by Corollary 1.6, with probability $1 - \delta$:

$$E_P(\chi(f(x, y))) \leqslant E_P(g(f(x, y))) \leqslant$$
$$\leqslant \tilde{E}_S(g(f(x, y))) + 2\tilde{\mathcal{R}}_l(g \circ \mathcal{F}_2) + 3\sqrt{\frac{\ln(2/\delta)}{2l}}. \quad (2.54)$$

By definition of the margin slack variable (2.52):

$$g(-y_i f(x_i)) \leqslant \xi_i/\gamma$$

for $1 \leqslant i \leqslant l$.

Let us bound the empirical Rademacher average of the class $\mathcal{F}_2$:

$$\tilde{\mathcal{R}}_l(\mathcal{F}_2) = E_\sigma \left(\sup_{f \in \mathcal{F}_2} \frac{1}{l} \sum_{i=1}^{l} \sigma_i f(x_i, y_i) \right) =$$
$$= E_\sigma \left(\sup_{f \in \mathcal{F}_1} \frac{1}{l} \sum_{i=1}^{l} \sigma_i y_i f(x_i) \right) =$$
$$= E_\sigma \left(\sup_{f \in \mathcal{F}_1} \frac{1}{l} \sum_{i=1}^{l} \sigma_i f(x_i) \right) =$$
$$= \tilde{\mathcal{R}}_l(\mathcal{F}_1) \leqslant \frac{1}{l} \sqrt{tr(K)}.$$

Since the function g is Lipschitz continuous with the constant $L = 1/\gamma$, we have by Theorem 1.12 $\tilde{\mathcal{R}}_l(g \circ \mathcal{F}_2) \leqslant \tilde{\mathcal{R}}_l(\mathcal{F}_2)/\gamma = \tilde{\mathcal{R}}_l(\mathcal{F}_1)/\gamma$. By definition for any $f \in \mathcal{F}_2$

$$\tilde{E}_S(g \circ f)) = \frac{1}{l}\sum_{i=1}^{l} g(-y_i f(\bar{x}_i)) \leqslant \frac{1}{l\gamma}\sum_{i=1}^{l}\xi_i.$$

By the inequalities (2.54) and (2.51) of Theorem 2.7, with probability $1-\delta$:

$$E_P(\chi(f(x,y))) \leqslant \frac{1}{l\gamma}\sum_{i=1}^{l}\xi_i + \frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\delta)}{2l}}. \qquad (2.55)$$

Theorem is proved. $\triangle$

In particular, in the case, where a function $f(x)$ separates a sample S without errors, the following upper bound is valid:

Corollary 2.1. *Assume that a function $f(\bar{x})$ separates a sample S without errors, and all assumptions used in Theorem 2.8 hold. Then for any $\delta > 0$, with probability $1-\delta$,*

$$P^l\{y \neq \operatorname{sign}(f(x))\} \leqslant \frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2l}}.$$

In the example considered above, $f(\bar{x}) = (\bar{w}\cdot\phi(\bar{x}))$, where $\bar{x} \in \mathcal{R}^n$, $\bar{w} \in \mathcal{R}^N$.

Unlike the bound (1.30) and (2.33) obtained in the theory of fat-shattering dimension the bound (2.53) has best constants and does not require prior knowledge of the radius of the ball containing vectors of the training sample.

The bound (2.53) is worse than a similar estimate obtained using fat-shattering dimension. Let $\|\bar{x}_i\| \leqslant R$ for all $1 \leqslant i \leqslant l$. For small valies, the order of the variable:

$$\frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} \leqslant \frac{2}{l\gamma}\sqrt{lR^2} = 2\sqrt{\frac{R^2}{l\gamma^2}}$$

is much more than the order of the leading term of the bound (1.30) of Theorem 1.9 and the order of the leading term of the bound (2.33)

of Theorem 2.4, which are of order $O\left(\frac{R^2}{l\gamma^2}\right)$. The bounds (1.30) and (2.33) have been obtained in the theory of the fat-shattering dimension.

2.8. Multidimensional regression

2.8.1. Linear regression

Let$S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$ be a training sample, where $\bar{x}_i \in \mathcal{R}^n$, $y_i \in \mathcal{R}$ for $i = 1, \ \ldots, \ l$.

Linear regression problem consists in searching for a linear function

$$f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b$$

interpolating the elements of a sample S with with the greatest accuracy. Geometrically, this function is a hyperplane that approximates values y_i on arguments $\bar{x}_i \ \ i = 1, \ \ldots, \ l$.

This problem was solved by Gauss and Legendre in the XVIII century by minimizing the sum of squared differences of values $f(\bar{x}_i)$ and y_i for $i = 1, \ \ldots, \ l$. The generalization theory for this method is well presented in mathematical statistics for linear models generating data with Gaussian random noise.

In that follows any vector $\bar{x}$ will be represented as a matrix of dimension $(n \times 1)$ or as a column vector

$$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}.$$

We also use a transposed form of this vector – a row vector $\bar{x}' = (x_1, \ \ldots, \ x_n)$.

The product of two matrices A and B is denoted AB without a point between them. We often identify the dot product of vectors

$(\bar{x} \cdot \bar{z})$ and the matrix $\bar{x}'\bar{z}$:

$$\bar{x}'\bar{z} = (x_1, \ \ldots, \ x_n) \begin{pmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{pmatrix} = (x_1 z_1 + \ldots x_n z_n)$$

of dimension (1×1) with a single element equal to this scalar product.

By the method of least squares, we minimize the *square loss function*:

$$L(\bar{w}, b) = \sum_{i=1}^{l} (y_i - (\bar{w} \cdot \bar{x}_i) - b)^2. \tag{2.56}$$

Let us denote by $\tilde{w}$ an extended column vector of weight coefficients and the constant term:

$$\tilde{w} = \begin{pmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \\ b \end{pmatrix}.$$

Similarly, denote by $\tilde{x}$ an extended column vector of variables:

$$\tilde{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \\ 1 \end{pmatrix}.$$

In the new extended variables, the regression function has a homogeneous form:

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}). \tag{2.57}$$

Consider a matrix of dimension $(l \times (n+1))$, whose rows are expanded row vectors $\tilde{x}_i' = (\bar{x}_i', 1)$:

$$\tilde{X} = \begin{pmatrix} \tilde{x}_1' \\ \tilde{x}_2' \\ \cdot \\ \cdot \\ \cdot \\ \tilde{x}_l' \end{pmatrix} = \begin{pmatrix} x_{11}, & \ldots, & x_{1n}, 1 \\ x_{21}, & \ldots, & x_{2n}, 1 \\ & & \cdot \\ & & \cdot \\ & & \cdot \\ x_{l1}, & \ldots, & x_{ln}, 1 \end{pmatrix}.$$

Define an l-dimensional column vector

$$\bar{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_l \end{pmatrix}.$$

Differences $|f(\bar{x}_i) - y_i|$ (and also $y_i - f(\bar{x}_i)$ and $f(\bar{x}_i) - y_i$) are called *residuals*. The row vector of residuals is of the form $\bar{y} - (\tilde{X} \cdot \tilde{w})$. Then the functional (2.56) can be represented as a squared norm of the row vector:

$$L(\tilde{w}) = \|\tilde{X}\tilde{w} - \bar{y}\|^2 = (\bar{y} - \tilde{X}\tilde{w})'(\bar{y} - \tilde{X}\tilde{w}).$$

In what follows denote by A' the transpose of a matrix A.

Now the regression problem can be written as the problem of minimizing the squared norm of the vector of residuals:

$$L(\tilde{w}) = \|\tilde{X}\tilde{w} - \bar{y}\|^2 \to \min. \tag{2.58}$$

Geometrically, this can be interpreted just as the search for a projection of the vector $\bar{y}$ with the smallest length on the hyperplane generated by column vectors of the matrix $\tilde{X}$.

For finding the minimum equate the partial derivatives of this functional with respect to $w_1, \ \ldots, \ w_n, b$ to zero and obtain a system of $n+1$ equations:

$$\frac{\partial L(\tilde{w})}{\partial \tilde{w}} = -2\tilde{X}'\bar{y} + 2\tilde{X}'\tilde{X}\tilde{w} = \bar{0}.$$

Transform this system to the form:

$$\tilde{X}'\tilde{X}\tilde{w} = \tilde{X}'\bar{y}.$$

If the matrix $X'\tilde{X}$ is invertible, we obtain the solution of this system:

$$\tilde{w} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\bar{y}.$$

2.8.2. Ridge regression

Ridge regression – a method supporting the numerical stability, was discovered by Hoerl and Kennard.

Recall that in order to get rid of the constant term in the regression equation, we consider the problem of regression with the extended column vector $\tilde{w}$ of weight coefficients and the constant term

$$\tilde{w} = \begin{pmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \\ b \end{pmatrix},$$

and also, $\tilde{x}$ – the extended column vector of variables

$$\tilde{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \\ 1 \end{pmatrix}.$$

In these new variables, the regression function has a homogeneous form without a constant term

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}).$$

Let us conside the following loss function:

$$L(\tilde{w}) = \lambda(\tilde{w} \cdot \tilde{w}) + \sum_{i=1}^{l}(y_i - (\tilde{w} \cdot \tilde{x}_i))^2 = \\ = \lambda\|\tilde{w}\|^2 + \|\tilde{X}\tilde{w} - \bar{y}\|^2. \qquad (2.59)$$

The parameter λ controls a balance between the square loss and the norm of the weight vector.

Ridge regression in the primal form

To find the extremum equate the to zero the partial derivatives: $L(\tilde{w})\ \ w_i$:

$$\lambda\tilde{w} - \sum_{i=1}^{l}((y_i - (\tilde{w} \cdot \tilde{x}_i))\tilde{x}_i = \tilde{0},$$

$i = 1,\ \ldots,\ n+1$. In matrix form, this equation is

$$\lambda\tilde{w} - \tilde{X}'\tilde{y} + \tilde{X}'\tilde{X}\tilde{w} = \tilde{0}.$$

The solution is written in matrix form:

$$\tilde{w} = (\lambda I + \tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y},$$

where I is the unit matrix.

The matrices $\tilde{X}'\tilde{X}$, I and $\lambda I + \tilde{X}'\tilde{X}$ have the size $(n+1)\times(n+1)$. The matrix $\tilde{X}'\tilde{X}$ is positive definite, ie

$$\tilde{z}'(\tilde{X}'\tilde{X})\tilde{z} \geqslant 0$$

for each vector $\tilde{z}$. This follows from the equality:

$$\tilde{z}'(\tilde{X}'\tilde{X})\tilde{z} = (\tilde{X}\tilde{z})'(\tilde{X}\tilde{z}) = \|\tilde{X}\tilde{z}\|^2 \geqslant 0.$$

Once added to the matrix $\tilde{X}'\tilde{X}$ the matrix λI, where $\lambda > 0$, a new matrix becomes strictly positive definite:

$$\tilde{z}'(\lambda I + \tilde{X}'\tilde{X})\tilde{z} = \lambda\|\tilde{z}\|^2 + \|\tilde{X}\tilde{z}\|^2 > 0$$

for $\tilde{z} \neq \tilde{0}$. It is well known that any positive definite matrix is invertible. Therefore, a solution of the problem of ridge regression is always exists for $\lambda > 0$.

When $\lambda = 0$ the matrix $\tilde{X}'\tilde{X}$ can be non-invertible. In this case, solution of regression problem is not unique. Therefore, the problem of ridge regression with $\lambda > 0$ is numerically much simpler than the problem of simple regression. Also, the parameter λ plays the role of the penalty for a large norm of the weight vector $\tilde{w}$.

If λ is approaching zero, the matrix $\lambda I + \tilde{X}'\tilde{X}$ can become closer to a non-invertible matrix. In this case, the inversion algorithm of this matrix is becoming increasingly unstable. The large values λ make the process of computing the inverse matrix more stable.

On the other hand, for large values of λ the matrix λI starts to predominate over the matrix $\tilde{X}'\tilde{X}$ and, so, regression residuals become larger and the regression predictor loses its predictive capabilities. So the value of λ must be of the same order as the elements of the matrix $\tilde{X}'\tilde{X}$.

The dual form of the problem of ridge regression and its generalization to the nonlinear case will be considered in Section 2.9.2.

2.9. Support vector regression

2.9.1. Solution of the problem of regression with SVM

Support vector machines are also used for solving the problem of regression. In this case, as well as in the problem of classification, nonlinear separating functions correspond to linear separating functions in a feature space defined by a kernel.

By a linear ϵ-insensative loss function we mean a function

$$L^{\epsilon}(\bar{x}, y, f) = |y - f(\bar{x})|_{\epsilon} = \max\{0, |y - f(\bar{x})| - \epsilon\}, \tag{2.60}$$

where f is an arbitrary function of type $\mathcal{R}^n \to \mathcal{R}$.

Let $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ be a vector of the margin slack variables, where $\xi_i = L^{\epsilon}(\bar{x}_i, y_i, f)$, $i = 1, \ \ldots, \ l$.

Similarly, ϵ-insensative quadratic loss function is defined as

$$L_2^{\epsilon}(\bar{x}, y, f) = (|y - f(\bar{x})|_{\epsilon})^2. \tag{2.61}$$

Quadratic ϵ-insensitive loss

In this case we have to minimize the quantity:

$$R^2\|\bar{w}\|^2 + \sum_{i=1}^{l} L_2^{\epsilon}(\bar{x}_i, y_i, f),$$

where $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b$.

We will minimize the loss function at once for all possible values of γ and for all possible values of $\theta = \epsilon + \gamma$ for fixed $\epsilon > 0$. To do this we introduce in the optimization problem the variables ξ_i and $\hat{\xi}_i$, which control the deviations of residuals by a value ϵ in greater or down direction from a fixed boundary. The parameter C controls a balance between the complexity of a regression hypothesis and sums of corresponding quadratic residuals.

The primal optimization problem for the square loss function (2.61) and for fixed values of the parameters C and ϵ is formulated as follows:

$$\begin{aligned} \|\bar{w}\|^2 + C\sum_{i=1}^{l}(\xi_i^2 + \hat{\xi}_i^2) &\to \min \\ \text{subject to } ((\bar{w} \cdot \bar{x}_i) + b) - y_i &\leqslant \epsilon + \xi_i,\ i = 1,\ \ldots,\ l, \\ y_i - ((\bar{w} \cdot \bar{x}_i) + b) &\leqslant \epsilon + \hat{\xi}_i,\ i = 1,\ \ldots,\ l, \\ \xi_i, \hat{\xi}_i &\geqslant 0,\ i = 1,\ \ldots,\ l. \end{aligned} \tag{2.62}$$

In practical applications, the parameter C can be defined using a procedure of exhaustive search.

The Lagrangian of the primal problem is:

$$\begin{aligned} L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha}) = |\bar{w}|^2 + C\sum_{i=1}^{l}(\xi_i^2 + \hat{\xi}_i^2) + \\ + \sum_{i=1}^{l}\alpha_i((\bar{w} \cdot \bar{x}_i) + b - y_i - \epsilon - \xi_i) + \\ + \sum_{i=1}^{l}\hat{\alpha}_i(y_i - (\bar{w} \cdot \bar{x}_i) - b - \epsilon - \hat{\xi}_i), \end{aligned}$$

where $\bar{\alpha} = (\alpha_1,\ \ldots,\ \alpha_l)$ and $\hat{\alpha} = (\hat{\alpha}_1,\ \ldots,\ \hat{\alpha}_l)$.

Note that just like in the classification problem, the conditions $\xi_i \geqslant 0$ and $\hat{\xi}_i \geqslant 0$ may be omitted since any solution where $\xi_i < 0$ or $\hat{\xi}_i < 0$ can be transformed into a solution where $\xi_i = 0$ or $\hat{\xi}_i = 0$.

To find the minimum equate the partial derivatives of the Lagrangian to zero:

$$\begin{aligned}
\frac{\partial L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha})}{\partial \bar{w}} &= \bar{0}, \\
\frac{\partial L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha})}{\partial b} &= 0, \\
\frac{\partial L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha})}{\partial \bar{\xi}} &= \bar{0}, \\
\frac{\partial L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha})}{\partial \hat{\xi}} &= \bar{0}.
\end{aligned}$$

From the first equation, we obtain an expression for the weight vector:

$$\bar{w} = \frac{1}{2} \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) \bar{x}_i. \tag{2.63}$$

Note that for any valid solution of (2.62), we have $\xi_i \hat{\xi}_i = 0$ for all i Therefore, for the dual problem $\alpha_i \hat{\alpha}_i = 0$.

The corresponding dual problem is formulated as follows:

$$\begin{gathered}
\sum_{i=1}^{l} y_i (\hat{\alpha}_i - \alpha_i) - \epsilon \sum_{i=1}^{l} (\hat{\alpha}_i + \alpha_i) - \\
-\frac{1}{2} \sum_{i,j=1}^{l} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)((\bar{x}_i \cdot \bar{x}_j) + \frac{1}{C} \delta_{ij}) \to \max \\
\text{subject to } \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) = 0, \\
\hat{\alpha}_i \geqslant 0, \alpha_i \geqslant 0, \ i = 1, \ \ldots, \ l,
\end{gathered} \tag{2.64}$$

where $\delta_{ij} = 1$ if and only if $i = j$.

The Karush–Kuhn–Tucker conditions are:

$$\begin{aligned}
&\alpha_i((\bar{w}\cdot\bar{x}_i)+b-y_i-\epsilon-\xi_i)=0,\ i=1,\ \ldots,\ l,\\
&\hat{\alpha}_i(y_i-(\bar{w}\cdot\bar{x}_i)-b-\epsilon-\hat{\xi}_i)=0,\ i=1,\ \ldots,\ l,\\
&\xi_i\hat{\xi}_i=0,\ \alpha_i\hat{\alpha}_i=0,\ i=1,\ \ldots,\ l. \qquad (2.65)
\end{aligned}$$

Denote $\beta_i=\hat{\alpha}_i-\alpha_i$, $i=1,\ \ldots,\ l$. Then using equalities $\alpha_i\hat{\alpha}_i=0$ for all i, the dual problem (2.64) is similar to a dual problem for classification:

$$\begin{aligned}
&\sum_{i=1}^{l}y_i\beta_i-\epsilon\sum_{i=1}^{l}|\beta_i|-\\
&-\frac{1}{2}\sum_{i,j=1}^{l}\beta_i\beta_j((\bar{x}_i\cdot\bar{x}_j)+\frac{1}{C}\delta_{ij})\to\max\\
&\text{subject to}\sum_{i=1}^{l}\beta_i=0,\ i=1,\ \ldots,\ l. \qquad (2.66)
\end{aligned}$$

It follows from the Karush–Kuhn–Tucker conditions (2.65) that $\alpha_i=\hat{\alpha}_i=0$ for all vectors $\bar{x}_i$ fallen into a layer of width ϵ around the regression hyperplane. Therefore, in the sum (2.63), the corresponding terms are absent. The number of support vectors decreases, and the the dual maximization problem is simplified. Support vectors are those vectors $\bar{x}_i$ for which $(\bar{w}\cdot\bar{x}_i)+b\leqslant y_i-\epsilon$ or $(\bar{w}\cdot\bar{x}_i)+b\geqslant y_i+\epsilon$.

Kernel SVM regression

Since the sample vectors are used in the optimization problem only through inner products, we can replace them by their images in a feature space and move on to a kernel version.

For simplicity, replace $\beta_i=\hat{\alpha}_i-\alpha_i$ on α_i. So α_i has a different meaning in this subsection than in the past. The kernel version of the regression problem is formulated in the form:

Theorem 2.9. *Let $S=((\bar{x}_1,y_1),\ \ldots,\ (\bar{x}_l,y_l))$ be a training sample, where $\bar{x}_i\in\mathcal{X}$ and $y_i\in\mathcal{R}$. Let also, a kernel $K(\bar{x},\bar{z})$ defines a feature space and $\bar{\alpha}^*$ be a solution of the corresponding quadratic optimization*

problem:

$$W(\bar{\alpha}) = \sum_{i=1}^{l} y_i \alpha_i - \epsilon \sum_{i=1}^{l} |\alpha_i| -$$

$$-\frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j (K(\bar{x}_i, \bar{x}_j) + \frac{1}{C} \delta_{ij}) \to \max$$

$$\textit{subject to } \sum_{i=1}^{l} \alpha_i = 0, \ i = 1, \ \ldots, \ l. \tag{2.67}$$

Also, $f(\bar{x}) = \sum\limits_{i=1}^{l} \alpha_i K(\bar{x}_i, \bar{x}) + b^*$*, where* b^* *is such that* $f(\bar{x}_i) - y_i = -\epsilon - \alpha_i / C$ *for any* i *with* $\alpha_i > 0$.

Then the function $f(\bar{x})$ *is equivalent to a hyperplane in the feature space defined by the kernel* $K(\bar{x}_i, \bar{x})$ *which solves the optimization problem (2.62).*

Linear ϵ-insensitive loss function

The regression problem for the linear ϵ-insensitive loss function (2.60) is considered similarly. We will minimize the function:

$$\frac{1}{2} \|\bar{w}\|^2 + C \sum_{i=1}^{l} L^{\epsilon}(\bar{x}_i, y_i, f),$$

where $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b$, C is a parameter controlling a balance between the complexity of a regression hypothesis and a sum of linear residuals.

The primal optimization problem in case of ϵ-insensitive loss function (2.60) for given parameters C and ϵ is formulated as follows:

$$\frac{1}{2} \|\bar{w}\|^2 + C \sum_{i=1}^{l} (\xi_i + \hat{\xi}_i) \to \min$$

$$\text{subject to } ((\bar{w} \cdot \bar{x}_i) + b) - y_i \leqslant \epsilon + \xi_i, \ i = 1, \ \ldots, \ l,$$

$$y_i - ((\bar{w} \cdot \bar{x}_i) + b) \leqslant \epsilon + \hat{\xi}_i, \ i = 1, \ \ldots, \ l,$$

$$\xi_i, \hat{\xi}_i \geqslant 0, \ i = 1, \ \ldots, \ l. \tag{2.68}$$

The Lagrangian of the primal problem is:

$$L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha}) = \|\bar{w}\|^2 + C\sum_{i=1}^{l}(\xi_i + \hat{\xi}_i) +$$
$$+\sum_{i=1}^{l}\alpha_i((\bar{w}\cdot\bar{x}_i) + b - y_i - \epsilon - \xi_i) +$$
$$+\sum_{i=1}^{l}\hat{\alpha}_i(y_i - (\bar{w}\cdot\bar{x}_i) - b - \epsilon - \hat{\xi}_i).$$

The corresponding to (2.68) dual problem is formulated as follows:

$$\sum_{i=1}^{l} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon\sum_{i=1}^{l}(\hat{\alpha}_i + \alpha_i) -$$
$$-\frac{1}{2}\sum_{i,j=1}^{l}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)(\bar{x}_i\cdot\bar{x}_j) \to \max$$
$$\text{subject to } \sum_{i=1}^{l}(\hat{\alpha}_i - \alpha_i) = 0,$$
$$\hat{\alpha}_i \geqslant 0, \alpha_i \geqslant 0, \tag{2.69}$$
$$0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C,\ i = 1,\ \ldots,\ l. \tag{2.70}$$

The corresponding Karush–Kuhn–Tucker conditions are:

$$\alpha_i((\bar{w}\cdot\bar{x}_i) + b - y_i - \epsilon - \xi_i) = 0,\ i = 1,\ \ldots,\ l,$$
$$\hat{\alpha}_i(y_i - (\bar{w}\cdot\bar{x}_i) - b - \epsilon - \hat{\xi}_i) = 0,\ i = 1,\ \ldots,\ l,$$
$$(\alpha_i - C)\xi_i = 0,\ (\hat{\alpha}_i - C)\hat{\xi}_i = 0, \tag{2.71}$$
$$\xi_i\hat{\xi}_i = 0,\ \alpha_i\hat{\alpha}_i = 0,\ i = 1,\ \ldots,\ l. \tag{2.72}$$

Support vectors are those vectors $\bar{x}_i$ for which $\alpha_i > 0$ or $\hat{\alpha}_i > 0$. If the point y_i is fallen outside the layer of width ϵ located around the optimal hyperplane then $\alpha_i = C$ or $\hat{\alpha}_i = C$.

It holds $0 < \alpha_i < C$ or $0 < \hat{\alpha}_i < C$ only for vectors $\bar{x}_i$ such that the corresponding label y_i is located on the boundary of the layer.

Vectors $\bar{x}_i$, for which the values y_i are located inside the layer are certainly not support vectors, and for them $\alpha_i = 0$ and $\hat{\alpha}_i = 0$, since in this case the following inequalities hold

$$(\bar{w} \cdot \bar{x}_i) + b + \epsilon < y_i, \qquad \xi_i > 0,$$

and

$$(\bar{w} \cdot \bar{x}_i) + b - \epsilon < y_i, \qquad \hat{\xi}_i > 0.$$

The weight vector is a linear combination of support vectors:

$$\bar{w} = \frac{1}{2} \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) \bar{x}_i.$$

It holds $\alpha_i \hat{\alpha}_i = 0$.

The regression function is:

$$f(\bar{x}) = \sum_{i=1}^{l} \beta_i (\bar{x}_i \cdot \bar{x}) + b^*,$$

where $\beta_i = \hat{\alpha}_i - \alpha_i$.

The dual problem for the kernel version is formulated:

$$\sum_{i=1}^{l} y_i (\hat{\alpha}_i - \alpha_i) - \epsilon \sum_{i=1}^{l} (\hat{\alpha}_i + \alpha_i) -$$

$$-\frac{1}{2} \sum_{i,j=1}^{l} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j) K(\bar{x}_i, \bar{x}_j) \to \max$$

$$\text{subject to } \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) = 0,$$

$$\hat{\alpha}_i \geqslant 0, \alpha_i \geqslant 0,$$

$$0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C, \ i = 1, \ \ldots, \ l.$$

The regression function for the kernel version:

$$f(\bar{x}) = \sum_{i=1}^{l} \beta_i K(\bar{x}_i, \bar{x}) + b^*,$$

where $\beta_i = \hat{\alpha}_i - \alpha_i$.

2.9.2. Ridge regression in the dual form

Ridge regression can be represented as a special case of support vector regression with ϵ-insensitive square loss function (2.61), where $\epsilon = 0$.

We illustrate the solution of this problem as a special case of support vector regression irrespective of the results of Section 2.8.2.

Consider the margin slack variables defined in Section 2.8.1; we will use the same notations. Then the regression function with extended variables $\tilde{x}$ has a homogeneous form:

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}).$$

The primal minimization problem is formulated as follows:

$$\lambda \|\tilde{w}\|^2 + \sum_{i=1}^{l} \xi_i^2 \to \min$$
$$\text{subject to } y_i - (\tilde{w} \cdot \tilde{x}_i) = \xi_i,\ i = 1,\ \ldots,\ l.$$

In this case, the Lagrangian has the form:

$$L(\tilde{w}, \bar{\xi}, \bar{\alpha}) = \lambda \|\tilde{w}\|^2 + \sum_{i=1}^{l} \xi_i^2 + \sum_{i=1}^{l} \alpha_i (y_i - (\tilde{w} \cdot \tilde{x}_i) - \xi_i). \quad (2.73)$$

Equating to zero the partial derivatives of the Lagrangian (2.73) by w_j and ξ_j, we obtain:

$$\frac{\partial L(\tilde{w}, \bar{\xi}, \bar{\alpha})}{\partial \tilde{w}} = 2\lambda \tilde{w} - \sum_{i=1}^{l} \alpha_i \tilde{x}_i = 0,$$
$$2\xi_i - \alpha_i = 0$$

for $i = 1,\ \ldots,\ l$. Let us express the weight vector of regression functions and the margin slack variables through variables of the dual problem:

$$\tilde{w} = \frac{1}{2\lambda} \sum_{i=1}^{l} \alpha_i \tilde{x}_i,$$
$$\xi_i = \frac{\alpha_i}{2}.$$

At first, calculate:

$$\lambda(\tilde{w}\cdot\tilde{w}) = \frac{1}{4\lambda}\sum_{i,j=1}^{l}\alpha_i\alpha_j(\tilde{x}_i\cdot\tilde{x}_j),$$

$$\sum_{i=1}^{l}\alpha_i(\tilde{w}\cdot\tilde{x}_i) = \frac{1}{2\lambda}\sum_{i,j=1}^{l}\alpha_i\alpha_j(\tilde{x}_i\cdot\tilde{x}_j).$$

Substituting these expressions in (2.73), we obtain the dual problem

$$W(\bar{\alpha}) = \sum_{i=1}^{l} y_i\alpha_i - \frac{1}{4\lambda}\sum_{i,j=1}^{l}\alpha_i\alpha_j(\tilde{x}_i\cdot\tilde{x}_j) - \\ -\frac{1}{4}\sum_{i=1}^{l}\alpha_i^2 \to \max. \qquad (2.74)$$

This problem can be rewritten in the vector form:

$$W(\bar{\alpha}) = \bar{y}'\bar{\alpha} - \frac{1}{4\lambda}\bar{\alpha}'K\bar{\alpha} - \frac{1}{4}\bar{\alpha}'\bar{\alpha} \to \max,$$

where K is the Gram matrix; its elements are pairwise dot products of vectors $K_{i,j} = (\tilde{x}_i\cdot\tilde{x}_j)$.

Equating to zero the partial derivatives of $W(\bar{\alpha})$ (defined by (2.74)) by α_i, we obtain the system of equations in the vector form:

$$-\frac{1}{2\lambda}K\bar{\alpha} - \frac{1}{2}\bar{\alpha} + \bar{y} = \bar{0}.$$

The solution of this equation in the vector form is written as:

$$\bar{\alpha} = 2\lambda(K+\lambda I)^{-1}\bar{y}. \qquad (2.75)$$

Therefore, we have obtained the regression equation in the dual form.

We represent the dot product of the extended weight vector and the vector the extended variables:

$$(\tilde{w}\cdot\tilde{x}) = \frac{1}{2\lambda}\sum_{i=1}^{l}\alpha_i(\tilde{x}_i\cdot\tilde{x}) = \frac{1}{2\lambda}(\bar{\alpha}'\cdot\bar{k}),$$

where $\bar{\alpha} = (\alpha_1, \ \ldots, \ \alpha_l)$, $\bar{k} = (k_1, \ \ldots, \ k_l)$ for $k_i = (\tilde{x}_i \cdot \tilde{x})$.

Since the matrix K is symmetric, $K' = K$. Using identity $(AB)' = B'A'$ and (2.75), we obtain:

$$\bar{\alpha}' = 2\lambda \bar{y}'(K + \lambda I)^{-1}.$$

Then the regression function has the form:

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}) = \bar{y}'(K + \lambda I)^{-1}\bar{k}. \tag{2.76}$$

Note one shortcoming of this production. Since $\epsilon = \theta - \gamma = 0$, the number of parameters α_i is equal to l and the size of the matrix $K + \lambda I$ is equal to $l \times l$. By this reason, we cannot use for learning too much sample.

In the case of large sample the data can be separated into clusters and a regression hyperplane can be constructed for each cluster separately.

Non-linear kernel ridge regression

Dual form of the regression serves as a basis for considering non-linear regression defined by a kernel $K(\tilde{x}, \tilde{y})$.

We recall the scheme of transition to nonlinear regression in more detail. Consider a mapping $\tilde{x} \to \bar{\phi}(\tilde{x})$ of the input space to a feature space $\mathcal{R}^N$ of greater dimension. The dot product in $\mathcal{R}^N$ defines a kernel $K(\tilde{x}_i, \tilde{x}_j) = (\bar{\phi}(\tilde{x}_i) \cdot \bar{\phi}(\tilde{x}_j))$. The corresponding Gram matrix has the form:

$$K = \begin{pmatrix} K(\tilde{x}_1, \tilde{x}_1), & \ldots, & K(\tilde{x}_1, \tilde{x}_l) \\ K(\tilde{x}_2, \tilde{x}_1), & \ldots, & K(\tilde{x}_2, \tilde{x}_l) \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ K(\tilde{x}_l, \tilde{x}_1), & \ldots, & K(\tilde{x}_l, \tilde{x}_l) \end{pmatrix}.$$

Let $\bar{z}$ be the vector:

$$\bar{z} = \begin{pmatrix} K(\tilde{x}_1, \tilde{x}) \\ K(\tilde{x}_2, \tilde{x}) \\ \cdot \\ \cdot \\ \cdot \\ K(\tilde{x}_l, \tilde{x}) \end{pmatrix}.$$

Let $\bar{z} = (z_1, \ \ldots, \ z_l)$ and $z_i = K(\tilde{x}, \tilde{x}_i)$ for $i = 1, \ \ldots, \ l$. A non-linear hypersurface:

$$f(\tilde{x}) = \bar{y}'(\lambda I + K)^{-1}\bar{z} \tag{2.77}$$

is a preimage of a linear hyperplace (2.76) constructed in the feature space $\mathcal{R}^N$ by images $\bar{\phi}(\tilde{x}_1), \ \ldots, \ \bar{\phi}(\tilde{x}_l)$ of vectors of the training sample.

Probabilistic analogue of ridge regression with an arbitrary kernel is called *Krieging*. In probabilistic setting, vectors $\tilde{x}_1, \ \ldots, \ \tilde{x}_l$ are random variables with a given covariance function $R(\tilde{x}_i, \tilde{x}_j) = E(\tilde{x}_i \cdot \tilde{x}_j)$ known up to a small number of parameters.

2.10. Non-linear optimization

The main advantage of the support vector method is associated with the use of dual representation of the problem. The dual optimization problem is not only simplifies the boundary conditions for an optimization problem, but also provides the weights of a separating hyperplane (hypersurface) through support vectors. This representation does not depend on dimension of the input space. It can be considered as a method of compressing the information contained in a training set.

In this section we consider the direct and dual optimization problems and give their basic properties.

The primal optimization problem

Let the real valued functions $f(\bar{w})$, $g_i(\bar{w})$, and $h_i(\bar{w})$ with domain $\mathcal{R}^n$ be given, where $i = 1, \ \ldots, \ m$ and $\bar{w} \in \mathcal{R}^n$. The problem is to find an infimum:

$$\inf_{\bar{w}} f(\bar{w}) \text{ subject to constraints:}$$

$$g_i(\bar{w}) \leqslant 0, \, i = 1, \ \ldots, \ m, \tag{2.78}$$

$$h_i(\bar{w}) = 0, \, i = 1, \ \ldots, \ m, . \tag{2.79}$$

where $f(\bar{w})$ is called the objective function and (2.78) and (2.79) are called the equality and inequality constraints.

The last two constraints can be written in the vector form: $\bar{g}(\bar{w}) \leqslant \bar{0}$ and $\bar{h}(\bar{w}) = \bar{0}$. Let

$$\mathcal{R} = \{\bar{w} \in \mathcal{R}^n : \bar{g}(\bar{w}) \leqslant \bar{0}, \bar{h}(\bar{w}) = 0\}$$

be the feasible region.

A solution of the optimization problem is a vector $\bar{w}^*$ such that $\bar{w}^* \in \mathcal{R}$ and there exists no vector $\bar{w} \in \mathcal{R}^n$ for which $f(\bar{w}) < f(\bar{w}^*)$. Such a vector is also called a global minimum. A point $\bar{w}^*$ is called a local minimum of $f(\bar{w})$ if the same property holds in some neighborhood of $\bar{w}^*$.

If $f(\bar{w})$ is a quadratic function from coordinates of $\bar{w}$ and $\bar{g}$, $\bar{h}$ are linear functions then the optimization problem is called the quadratic optimization problem.

A real valued function f is called convex if for any $\bar{w}, \bar{u} \in \mathcal{R}^n$ and $0 \leqslant \lambda \leqslant 1$:

$$f(\lambda\bar{w} + (1-\lambda)\bar{u}) \leqslant \lambda f(\bar{w}) + (1-\lambda)f(\bar{u}).$$

Lagrangian theory is the optimization theory for the case, where there are only equality constraints $\bar{h}(\bar{w}) = \bar{0}$. The Lagrangian function is:

$$L(\bar{w}, \bar{\beta}) = f(\bar{w}) + \bar{\beta}\bar{h}(\bar{w}),$$

where the coefficients $\bar{\beta}$ are called the Lagrange multipliers.

A necessary condition for a minimum of $f(\bar{w})$ subject to $\bar{h}(\bar{w}) = \bar{0}$ is:

$$\frac{\partial L(\bar{w}, \bar{\beta})}{\partial \bar{w}} = \bar{0},$$
$$\frac{\partial L(\bar{w}, \bar{\beta})}{\partial \bar{\beta}} = \bar{0}.$$

The above conditions are also sufficient provided the function L convex by $\bar{w}$.

In the general case (2.79) the Lagrangian has the form:

$$L(\bar{w}, \bar{\alpha}, \bar{\beta}) = f(\bar{w}) + \sum_{i=1}^{m} \alpha_i g_i(\bar{w}) + \sum_{i=1}^{m} \beta_i h_i(\bar{w}) =$$
$$= f(\bar{w}) + \bar{\alpha}\bar{g}(\bar{w}) + \bar{\beta}\bar{h}(\bar{w}).$$

The dual optimization problem

The dual optimization problem is simpler than the primal optimization problem, since their constraints are simpler. Let

$$\Theta(\bar{\alpha}, \bar{\beta}) = \inf_{\bar{w}} L(\bar{\alpha}, \bar{\beta}, \bar{w}).$$

The dual optimization problem is to find a maximum:

$$\begin{aligned} &\max_{(\bar{\alpha},\bar{\beta})} \Theta(\bar{\alpha}, \bar{\beta}) \text{ subject to} \\ &\alpha_i \geqslant 0,\ i = 1,\ \ldots,\ m. \end{aligned} \tag{2.80}$$

The following is a weak duality theorem.

Theorem 2.10. *Let a vector $\bar{w}$ satisfies the conditions (2.78) and (2.79) of the primal optimization problem (in particular, it can be a solution of the primal problem) and $(\bar{\alpha}, \bar{\beta})$ be a solution of the dual problem (2.80). Then $f(\bar{w}) \geqslant \Theta(\bar{\alpha}, \bar{\beta})$.*

Proof. By definition:

$$\begin{aligned} \Theta(\bar{\alpha}, \bar{\beta}) = \inf_{\bar{u}} L(\bar{u}, \bar{\alpha}, \bar{\beta}) \leqslant L(\bar{w}, \bar{\alpha}, \bar{\beta}) = \\ = f(\bar{w}) + \bar{\alpha}\bar{g}(\bar{w}) + \bar{\beta}\bar{h}(\bar{w}) \leqslant f(\bar{w}). \end{aligned} \tag{2.81}$$

Here $\bar{\alpha}\bar{g}(\bar{w}) \leqslant 0$, since $\bar{\alpha} \geqslant \bar{0}$ and $\bar{g}(\bar{w}) \leqslant \bar{0}$, $\bar{h}(\bar{w}) = \bar{0}$. $\triangle$

This theorem immediately implies:

Corollary 2.2. *The value of the dual problem is upper bounded by the value of the primal problem:*

$$\sup\{\Theta(\bar{\alpha}, \bar{\beta}) : \bar{\alpha} \geqslant \bar{0}\} \leqslant \inf\{f(\bar{w}) : \bar{g}(\bar{w}) \leqslant \bar{0}, \bar{h}(\bar{w}) = \bar{0}\}.$$

Another consequence of this theorem gives a sufficient condition for that the values of solutions of the primal and dual problems coincide.

Corollary 2.3. *If $f(\bar{w}^*) = \Theta(\bar{\alpha}^*, \bar{\beta}^*)$, where $\bar{\alpha}^* \geqslant \bar{0}$, $\bar{g}(\bar{w}^*) \leqslant \bar{0}$, $\bar{h}(\bar{w}^*) = \bar{0}$, then $\bar{w}^*$ and $(\bar{\alpha}^*, \beta^*)$ are solutions of the primal and dual problems respectively. Also, $\bar{\alpha}^*\bar{g}(\bar{w}^*) = 0$.*

Proof. Since in the inequality (2.81) two extreme terms are equal, so it is an equality. In particular, $f(\bar{w}^*) = \inf\limits_{\bar{u}} L(\bar{u}, \bar{\alpha}^*, \bar{\beta}^*)$ and $\bar{\alpha}^* \bar{g}(\bar{w}^*) = 0$. $\triangle$

A sufficient condition for the existence of a solution of the primal and the dual problem is the existence of *a saddle point* of the Lagrangian. The saddle point $(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)$ satisfies the inequalities:

$$L(\bar{w}^*, \bar{\alpha}, \bar{\beta}) \leqslant L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*) \leqslant L(\bar{w}, \bar{\alpha}^*, \bar{\beta}^*)$$

for all $\bar{w}$, $\bar{\alpha}$, $\bar{\beta}$.

The strong duality theorem gives a sufficient condition for that the dual and primal problems have the same value for the optimization problems considered above.

Theorem 2.11. *Assume that the feasible set Ω is a convex subset of $\mathcal{R}^n$, the functions $\bar{h}$, $\bar{g}$ are affine (this means that $h_i(\bar{w})$, $g_i(\bar{w})$ has a form $A_i\bar{w} + \bar{b}_i$, where A_i is some matrix). Then solutions of the primal and the dual problems coincide.*

We now in a position to give the Kuhn–Tucker theorem giving conditions for an optimum solution to a general optimization problem.

Theorem 2.12. *Assume that a feasible set Ω is a convex subset of $\mathcal{R}^n$, the function f is convex, and the functions $\bar{h}$, $\bar{g}$ are affine.*

Then a vector $\bar{w}^*$ is a solution of the primal optimization problem:

$$\begin{aligned} \inf f(\bar{w}),\ \bar{w} \in \Omega,\ \textit{subject to} \\ \bar{g}(\bar{w}) \leqslant \bar{0}, \\ \bar{h}(\bar{w}) = \bar{0}, \end{aligned}$$

if and only if a pair $(\bar{\alpha}^*, \bar{\beta}^*)$ exists such that

$$\frac{\partial L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)}{\partial \bar{w}} = \bar{0},$$

$$\frac{\partial L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)}{\partial \bar{\beta}} = \bar{0}, \tag{2.82}$$

$$\alpha_i^* g_i(\bar{w}^*) = 0,\ i = 1,\ \ldots,\ m, \tag{2.83}$$

$$g_i(\bar{w}^*) \leqslant 0,\ i = 1,\ \ldots,\ m,$$

$$\alpha_i^* \geqslant 0,\ i = 1,\ \ldots,\ m.$$

Necessary conditions for maximum of the linear function $L(\bar{w}^*, \bar{\alpha}, \bar{\beta})$ by $\bar{\beta}$ are given by the conditions (2.82); these conditions are equivalent to the conditions: $h_i(\bar{w}^*) = 0$, $i = 1, \ \ldots, \ k$.

Necessary conditions for maximum of the linear function $L(\bar{w}^*, \bar{\alpha}, \bar{\beta})$ by α_i^* are contained in the conditions (2.83), since for $\alpha_i^* > 0$, any such condition is equivalent to the equality $g_i(\bar{w}^*) = 0$. The last equality is equivalent to the equality $\frac{\partial L(\bar{w}^*, \bar{\alpha}, \bar{\beta})}{\partial \alpha_i} = 0$. Also, $\alpha_i^* = 0$ at the maximum point of $L(\bar{w}^*, \bar{\alpha}, \bar{\beta})$.

The conditions (2.83) are called *the Karush–Kuhn–Tucker conditions*. They mean that if a solution the optimization problem is achieved at the boundary of the ith constraints then $\alpha_i^* \geqslant 0$, and $\alpha_i^* = 0$ otherwise.

Quadratic programming

The quadratic optimization problem is defined:

$$\begin{aligned} &\frac{1}{2}\bar{w}'Q\bar{w} - \bar{k}\bar{w} \to \min \\ &\quad \text{subject to } X\bar{w} \leqslant \bar{c}, \end{aligned} \tag{2.84}$$

where Q is an $n \times n$-positive definite matrix, $\bar{k}$ is an n-vector, $\bar{c}$ is an m-vector, $\bar{w}$ is an n-vector of vector of unknown variables, and X is an (m, n)-matrix.

Assume that these conditions define a non-emply set. Then the optimization problems can be rewritten:

$$\begin{aligned} \min_{\bar{w}} \left(\frac{1}{2}\bar{w}'Q\bar{w} - \bar{k}\bar{w} + \bar{\alpha}'(X\bar{w} - \bar{c}) \right) \to \max_{\bar{\alpha}} \\ \text{subject to } \bar{\alpha} \geqslant \bar{0}. \end{aligned} \tag{2.85}$$

A minimun in (2.85) by $\bar{w}$ is attained for

$$\bar{w} = Q^{-1}(\bar{k} - X'\bar{\alpha}).$$

We substitute this expression in (2.84) and obtain the dual problem:

$$\begin{aligned} -\frac{1}{2}\bar{\alpha}'P\bar{\alpha} - \bar{\alpha}'\bar{d} - \frac{1}{2}\bar{k}'Q\bar{k} \to \max \\ \text{subject to } \bar{\alpha} \geqslant \bar{0}, \end{aligned} \tag{2.86}$$

where $P = XQ^{-1}X'$, $\bar{d} = \bar{c} - XQ^{-1}\bar{k}$.

The dual problem is also quadratic, but their constraints are much simpler than the constraints of the primal problem.

2.11. Conformal predictions

Assume that an ordered sample $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$ be given, where $\bar{x}_i \in \mathcal{R}^n$ and $y_i \in \{-1, +1\}$ for $1 \leqslant i \leqslant l$. When solving the problem of classification using the separating hypersurface different examples of the sample are classified with a different degree of confidence.

Vovk and Gammerman [39] introduced a measure of non-conformity of an example $(\bar{x}_i, y_i)$. This non-conformity measure can be used to improve the performance of well-known prediction algorithms.

We define a measure of non-conformity for the problem of SVM classification. Recall the method of the SVM classification. The vectors $\bar{x}_i$ of the sample S are mapped to vectors $\bar{\phi}(\bar{x}_i)$ in a feature space defined by a kernel $K(\bar{x}, \bar{x}') = (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{x}'))$. After that, a separating hyperplane in the feature space is constructed, and the weight vectors of this hyperplane are expressed as a linear combination of images of support vectors:

$$\bar{w} = \sum_{i=1}^{l} y_i \alpha_i \bar{\phi}(\bar{x}_i),$$

where α_i are Lagrange multipliers obtained by solving the corresponding dual optimization problem.

The corresponding hypersurface in the initial space has the form:

$$f(\bar{x}) = \sum_{i=1}^{l} y_i \alpha_i K(\bar{x}_i, \bar{x}) + b.$$

Define *the non-conformity measure* of an example $(\bar{x}_i, y_i)$ be equal to the Lagrange multiplier α_i.

This definition is justified as follows. By Karush–Kuhn–Tucker conditions $\alpha_i = 0$ if $y_i((\bar{w} \cdot \bar{\phi}(\bar{x}_i)) + b) > 1$. Such vectors $\bar{\phi}(\bar{x}_i)$ are correctly classified and lies outside of the boundary hyperplanes.

Support vectors are those vectors $\bar{\phi}(\bar{x}_i)$, for which $y_i((\bar{w}\cdot\bar{\phi}(\bar{x}_i))+b) \leqslant 1$, also, $\alpha_i \geqslant 0$ and $\xi_i = \alpha_i/C$. These are the vectors $\bar{\phi}(\bar{x}_i)$ which are on the boundary hyperplanes or incorrectly classified by them, in this case $y_i((\bar{w}\cdot\bar{\phi}(\bar{x}_i))+b) < 1$. In case of linear norm $\alpha_i \leqslant C$, where C is a balance constant from the corresponding optimization problem. Therefore:

- The examples with $\alpha_i = 0$ are correctly classified and, so, they have the highest degree of conformity with the sample.
- The examples with positive values of α_i either lie on the boundary hyperplanes or are incorrectly classified and, therefore, the degree of conformity of the example the worse the greater the value α_i.

Define *p-value* of an example $(\bar{x}_i, y_i)$:

$$p_i = \frac{|\{j : \alpha_j \geqslant \alpha_i\}|}{l}.$$

By definition $0 \leqslant p_i \leqslant 1$. The small value of p_i means that the example $(\bar{x}_i, y_i)$ has one of the biggest non-conformity measure among the examples of the sample S.

We construct a meta-algorithm for conformal SVM classification using p-values.

Assume that a sample $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$ and unlabeled example $\bar{x}_{l+1}$ be given. We have to assign a label $y_{l+1} \in \{-1, +1\}$ to this vector.

Some level of confidence $\epsilon > 0$ also be given.

Meta-algorithm:

For each $y \in \{-1, +1\}$, solve the optimization problem of SVM classification using the extended sample:

$$S' = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l), (\bar{x}_{l+1}, y)),$$

find the values of Lagrange multipliers α_i, $1 \leqslant i \leqslant l+1$ and calculate p-value:

$$p(y) = \frac{|\{j : \alpha_j \geqslant \alpha_{l+1}\}|}{l+1}.$$

Output of the algorithm:

- if $p(y) < \epsilon$ for all y then algorithm does not output any result;
- if $p(y) \geqslant \epsilon$ for some y then output as a result the value of y, for which the quantity $p(y)$ takes its maximal value:

$$y_{l+1} = \arg\max_y p(y).$$

Such a procedure is justified by a probabilistic result which states that under certain probabilistic assumptions about the mechanism of generation of examples p-value satisfies the natural condition for a test:

$$P\{p_i \leqslant \epsilon\} \leqslant \epsilon,$$

where P is a measure on samples of α_i invariant with respect to their permutations (see Vovk et al. [39]).

A non-conformity measure is defined depending on specificity of a data model. Vovk et al. [39] constructed non-conformity measures for for the nearest neighbor algorithm, SVM, bootstrap, neural networks, decision trees, ridge regression and Bayes algorithm.

Consider an example of non-conformity measure for classification using the method of nearest neighbor. The idea of the method of k-nearest neighbors consist in the following. In order to predict the label of the new unlabeled object $\bar{x}$ we find k nearest by distance neighbors $\bar{x}_i$ of this object. In the classification problem, "a voting" method is used – we assign to the object $\bar{x}$ a label which occurs most frequently in the nearest k objects. In the regression problem, we can take the median of their labels.

We consider examples $(\bar{x}, y)$, where $\bar{x} \in \mathcal{R}^n$, $y \in D$, and D is a finite set. Assume that $\{\bar{x}_1, \ \ldots, \ \bar{x}_k\}$ be a set of all k nearest to $\bar{x}$ objects and $\{y_1, \ \ldots, \ y_k\}$ be their labels.

We define a non-conformity measure of an example $(\bar{x}, y)$, as the ratio of the minimal distance of the object $\bar{x}$ to the objects $\bar{x}_i$ with the same label $y_i = y$ to the minimum distance of the object $\bar{x}$ to the objects $\bar{x}_i$ with different labels $y_i \neq y$:

$$\alpha_{(\bar{x},y)} = \frac{\min_{1 \leqslant j \leqslant k, y_j = y} d(\bar{x}, \bar{x}_j)}{\min_{1 \leqslant j \leqslant k, y_j \neq y} d(\bar{x}, \bar{x}_j)}.$$

The distance $d(\bar{x}, \bar{x}')$ refers to the standard Euclidean distance between two vectors.

The greater the value $\alpha_{(\bar{x},y)}$ the closer the object $\bar{x}$ to other objects indicated by the labels different from y, ie, the greater the degree of non-conformity of the example $(\bar{x}, y)$.

2.12. Problems

1. Prove that the function $\rho(\bar{w})$ defined by (2.3) in Section 2.1 is concave. (*Hint*: We have to verify that

$$\rho(\lambda\bar{w} + (1-\lambda)\bar{u})) \geqslant \lambda\rho(\bar{w}) + (1-\lambda)\lambda(\bar{u}) \tag{2.87}$$

for all $0 \leqslant \lambda \leqslant 1$ and $\bar{w}$, $\bar{u}$ lying in the unit ball.

The following inequalities:

$$\begin{aligned} &\min_{i\in I}(f(i)+g(i)) \geqslant \min_{i\in I} f(i) + \min_{i\in I} g(i), \\ &\max_{i\in I}(f(i)+g(i)) \leqslant \max_{i\in I} f(i) + \max_{i\in I} g(i) \end{aligned} \tag{2.88}$$

hold for all functions f and g and sets I.

By definition

$$\rho(\bar{w}) = \frac{1}{2}(\min_{y_i=1}(\bar{w}\cdot\bar{x}_i) - \max_{y_i=-1}(\bar{w}\cdot\bar{x}_i)),$$

By (2.88), where $f(i) = (\bar{w}\cdot\bar{x}_i)$ and $g(i) = (\bar{u}\cdot\bar{x}_i)$, we have

$$\begin{aligned} &\min_{y_i=1}((\lambda\bar{w} + (1-\lambda)\bar{u})\cdot\bar{x}_i) = \\ &= \min_{y_i=1}(\lambda(\bar{w}\cdot\bar{x}_i) + (1-\lambda)(\bar{u}\cdot\bar{x}_i)) \geqslant \\ &\geqslant \lambda\min_{y_i=1}(\bar{w}\cdot\bar{x}_i) + (1-\lambda)\min_{y_i=1}(\bar{u}\cdot\bar{x}_i). \end{aligned}$$

A similar inequality holds for the maximum. Subtracting the corresponding inequalities, we obtain (2.87)).

2. Prove the rest part of Lemma 2.1.

3. Construct the mappings from $\mathcal{R}^n$ to a feature space and the corresponding polynomial kernels for polynomials of general form and

higher-order ($k = 3, 4, \ \ldots\ $), and for the appropriate classification functions of the form (2.24).

4. Prove that for any positive definite function $K(x, y)$ the Cauchy–Schwartz inequality holds:

$$K(x_1, x_2) \leqslant \sqrt{K(x_1, x_1)K(x_2, x_2)} \quad x_1, x_2 \in X. \tag{2.89}$$

Note: The eigenvalues of any positive definite (2×2) matrix $K(x_i, x_j)$ are non-negative. Therefore, the same holds for the determinant.

5. Prove that for any kernel $K(x, y)$:
(i) $K(x, x) \geqslant 0$ for all x.
(ii) If $K(x, x) = 0$ for all x then $K(x, y) = 0$ for all x and y.
Note that a function kernel $K(x, y)$ is non bilinear in general case.

6. Let $\mathcal{F}$ be a Hilbert space of functions on X such that any linear functional $f \to f(x)$ is continuous. By Riesz–Fisher theorem for each $x \in X$ there exists a function $K_x \in \mathcal{F}$ such that $f(x) = (K_x \cdot f)$ for all x. The reproducing kernel is determined $K(x, y) = (K_x \cdot K_y)$. Prove that the function $K(x, y) = (K_x \cdot K_y)$ is symmetric and positive definite.

7.Let $K_1(x, y), K_2(x, y), \ \ldots,$ be a kernel on a set X. Prove that the following combinations are also kernels:
(i) $\alpha_1 K_1(x, y) + \alpha_2 K_2(x, y)$, where $\alpha_1, \alpha_2 \geqslant 0$;
(ii) $K(x, y) = \lim\limits_{n\to\infty} K_n(x, y)$;
(iii) $K_1(x, y)K_2(x, y)$ (*Hint*: use a representation of any positive definite Gram matrix in the form $K = MM'$);
(iv) $K(A, B) = \sum\limits_{x\in A, y\in B} K(x, y)$, where A, B are finite subsets of X (this is a kernel on the set of all finite subsets of X).

Construct the corresponding mappings in the feature spaces.

8. Prove that in the optimization problem (2.67) a value of b^* does not depend on i.

9. Show that a dual problem corresponding to the primal problem (2.68) has a form (2.70). Prove relations (2.72) for the dual problem.

10. Prove that the Gram matrix $K_{i,j} = (\tilde{x}_i \cdot \tilde{x}_j)$ is invertible if and only if the vectors $\tilde{x}_1, \ \ldots, \ \tilde{x}_l$ are linear independent.

11. (i) Find the maximum volume of volume of of the parallelepiped for a given surface area.

(ii) Find the maximum of entropy $H(p_1, \ \ldots, \ p_n) = -\sum\limits_{i=1}^{n} p_i \ln p_i$ subject to $\sum p_i = 1$, $\sum c_i p_i = e$.

12. Perform all necessary calculations to obtain the solution (2.86) of the quadratic optimization problem.

13. Prove that given the class $\mathcal{F}$ of all linear (homogeneous) functions a set is γ-separable if and only if it is γ-separable (maybe for a different γ) on the same level, where $r = 0$.

14. Prove relations (2.66) for the dual regression problem.

2.13. Laboratory work

In this section we offer the basic laboratory work for solving the problem of classification using SVM.

Performance of the work includes the following procedures:

- Download the input data from the appropriate web-site. As a rule, the input data is a set of vectors of large dimension, for which the object classes are already specified.

- Divide the data into a training set and a test set. The object class is used in the training set for the training and in the test set for validation of the classification. Following the procedure of classification you have to count a proportion of correct answers.

- Calibrate (rescale) the original data if needed. Scaling of the data helps to avoid loss of accuracy due to too small or too large values of certain attributes. In particular, this is important when the Gaussian kernel is used. We recommend to normalize the numerical value of each feature so that it falls within the range of $[-1, 1]$ or $[0, 1]$.

- Choose the kernel best classifying the training set. Usually, the standard SVM software use the following kernels:

  1) the linear kernel $K(\bar{x}, \bar{y}) = (\bar{x} \cdot \bar{y})$,

  2) the polynomial kernel $K(\bar{x}, \bar{y}) = (\gamma(\bar{x} \cdot \bar{y}) + r)^d$, where $\gamma > 0$,

  3) the Gaussian kernel $K(\bar{x}, \bar{y}) = e^{-\frac{\|\bar{x} - \bar{y}\|^2}{\sigma^2}}$,

4) the sigmoid kernel $K(\bar{x}, \bar{y}) = tanh(\gamma(\bar{x} \cdot \bar{y}) + r)$.

We recommend to choose the Gaussian kernel $K(\bar{x}, \bar{y}) = e^{-\frac{\|\bar{x}-\bar{y}\|^2}{\sigma^2}}$ at first time.

- Cross-check the results in order to find the best values of the parameters C and γ. Note that it is not enough to find the parameter values that give the best accuracy only on the training set. To avoid the overfitting – to divide the sample set into two parts, to find the best values parameters at training on the first part, and to use the results of classification in the second part for the evaluation of the parameter estimation.

  There is a more complicated procedure of cross-validation, in which the training set is divided into N equal parts. Consistently choose one of the subsets, after that, train the classifier on the union of $N - 1$ remaining subsets and verify it on the selected subset. Fix the parameter values giving the greatest accuracy for one of these subsets.

  The parameters C and γ can also be selected using an exhaustive search on a discrete subset – lattice. The disadvantage of this the method is the big calculation time.

- Perform the constructed classifier on the test set. Compare the accuracy of classification for training and for the test sets.

There are a number of websites that contain SVM software relevant examples for experimental calculations. We mention some of them.

SVM software can be found at websites:
http://www.csie.ntu.edu.tw and www.support-vector.net

Website: http://archive.ics.uci.edu contains input data for solving classification and regression tasks.

Laboratory work 1

Conduct training and solve classification task for handwritten digits. Data for MATLAB can be found at website:
http://www.cs.toronto.edu

In particular, at this website, you can find data from the database USPS, containing digital images of the different spellings of handwritten digits.

Laboratory work 2

Provide training and classification with data of the following websites. Choose a data set to conduct training on SVM and testing on a test set.

LIBSVM library for support vector machines can be found at website: http://www.csie.ntu.edu.tw

The database for machine learning is located at website: http://www.csie.ntu.edu.tw

Laboratory work 3

Conduct training and classification on previous data with the perceptron and Rosenblatt's algorithm.

Part II

Prediction

Chapter 3

Universal prediction

3.1. Universal online forecasting

The following forecasting task is considered: for any n, a forecaster have to predict some information about a future outcome ω_n given past outcomes $\omega_1, \omega_2, \ldots, \omega_{n-1}$.

For simplicity, we consider in this section only binary outcomes $\omega_i \in \{0, 1\}$. In the measure-theoretic framework, we expect that the outcomes are generated by some probabilistic measure P and the conditional probabilities $p_n = P(\omega_n = 1 | \omega_1, \omega_2, \ldots, \omega_{n-1})$, $n = 1, 2, \ldots$, exist for all binary sequences $\omega_1, \omega_2, \ \ldots \ , \omega_{n-1}$ of length n. In this case, the forecaster must solve the classical statistical problem – reconstruction of the measure P given past observations.

Historically, the first universal prediction procedure was the Laplace rule. This procedure is based on the assumption that outcomes ω_i are generated by some i.i.d. source with the same probability p of generating 1 (and with probability $1 - p$ of generating 0). We do not know the true value of p, and we want to construct a forecasting procedure, which would be good enough for all p such that $0 \leqslant p \leqslant 1$.

Assume that we observe the outcomes $\omega^n = \omega_1, \ldots, \omega_n$, in which there are n_1 ones and n_2 zeros, where $n_1 + n_2 = n$. The probability of getting such a sequence of outcomes is $p^{n_1}(1 - p)^{n_2}$, where p is probability of generating 1. Since a true value of p is unknown, it is natural to consider the Bayesian mixture of oll such probabilities

with respect to the uniform measure:

$$P(\omega^n) = \int_0^1 p^{n_1}(1-p)^{n_2}dp.$$

The value of this integral is easy to calculate.

Lemma 3.1.

$$\int_0^1 p^{n_1}(1-p)^{n_2}dp = \frac{1}{(n+1)\binom{n}{n_1}}.$$

Proof. We prove this equality by backwards induction on n_1. For $n_1 = n$, it holds $\int_0^1 p^n dp = \frac{1}{(n+1)}$.

Assume that

$$\int_0^1 p^{n_1+1}(1-p)^{n_2-1}dp = \frac{1}{(n+1)\binom{n}{n_1+1}}.$$

Integrating by parts, we obtain

$$\int_0^1 p^{n_1}(1-p)^{n_2}dp = \frac{n-n_1}{n_1+1}\int_0^1 p^{n_1+1}(1-p)^{n_2-1}dp =$$
$$= \frac{n-n_1}{n_1+1}\frac{1}{(n+1)\binom{n}{n_1+1}} = \frac{1}{(n+1)\binom{n}{n_1}}.$$

Lemma is proved. $\triangle$

The conditional probability of the event $\omega_{n+1} = 1$ given past outcomes $\omega^n = \omega_1, \ \ldots, \ \omega_n$ is equal:

$$P\{\omega_{n+1} = 1|\omega^n\} = \frac{P(\omega^n 1)}{P(\omega^n)} = \frac{\frac{1}{(n+2)\binom{n+1}{n_1+1}}}{\frac{1}{(n+1)\binom{n}{n_1}}} = \frac{n_1+1}{n+2}.$$

Therefore, we obtain the *Laplace rule*:

$$P\{\omega_{n+1} = 1|\omega^n\} = \frac{n_1 + 1}{n + 2},$$
$$P\{\omega_{n+1} = 0|\omega^n\} = \frac{n_2 + 1}{n + 2}.$$

The performance of such forecasting procedure can be evaluated using some loss function. An example of such a loss function is the logarithmic loss function:

$$L_p(\omega^n) = -\ln(p^{n_1}(1-p)^{n_2}).$$

It is known from the information theory that this quantity coincides up to 1 with the average number of binary bits needed to encode the sequences ω^n, consisting of n_1 ones and n_2 zeros, and generated by a source with the given probability distribution.

For the Laplace rule:

$$L(\omega^n) = -\ln P(\omega^n) = -\ln \int\limits_0^1 p^{n_1}(1-p)^{n_2} dp.$$

Easy to verify that:

$$\sup_{0 \leqslant p \leqslant 1} p^{n_1}(1-p)^{n_2} = \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}.$$

Then for any sequence ω^n:

$$L(\omega^n) - \inf_{0 \leqslant p \leqslant 1} L_p(\omega^n) = \ln \frac{\sup\limits_{0 \leqslant p \leqslant 1} p^{n_1}(1-p)^{n_2}}{\int\limits_0^1 p^{n_1}(1-p)^{n_2} dp} =$$
$$= \ln \frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}}{\frac{1}{(n+1)\binom{n}{n_1}}} \leqslant \ln(n+1).$$

Thus, using the encoding probabilities, calculated by the Laplace rule, we will spend $\ln(n+1)$ extra bits as compared with the length

of the optimal code built on the basis of the true source generating outcomes ω_i.

Another, the more precise method of forecasting, was proposed by Krichevsky and Trofimov. We consider a Bayesian mixture of probabilities of all sequences of length n over all possible $0 < p < 1$ with the density $1/(\pi\sqrt{p(1-p)})$:

$$P(\omega^n) = \int\limits_0^1 \frac{p^{n_1}(1-p)^{n_2}}{\pi\sqrt{p(1-p)}} dp.$$

In this case, the conditional probability of generating 1 given past outcomes $\omega^n = \omega_1, \ \ldots, \ \omega_n$ is equal to

$$P(1|\omega^n) = \frac{n_1 + 1/2}{n+1}.$$

The following bound is valid:

$$\int\limits_0^1 \frac{p^{n_1}(1-p)^{n_2}}{\pi\sqrt{p(1-p)}} dp \geqslant \frac{1}{2\sqrt{n}} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}.$$

These statements are offered as the problems in Section 3.6 below.

From this we obtain the bound for the extra number of bits when encoding using the prediction by the method of Krichevsky and Trofimov:

$$L(\omega^n) - \inf_{0\leqslant p\leqslant 1} L_p(\omega^n) = \ln \frac{\sup\limits_{0\leqslant p\leqslant 1} p^{n_1}(1-p)^{n_2}}{\int_0^1 \frac{p^{n_1}(1-p)^{n_2}}{\pi\sqrt{p(1-p)}} dp} \leqslant$$

$$\leqslant \ln \frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}}{\frac{1}{2\sqrt{n}} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}} \leqslant \ln(2\sqrt{n}) = \frac{1}{2}\ln n + \ln 2.$$

In this bound the regret is asymptotically two times less than in the corresponding bound for the Laplace method.

3.2. Asymptotic calibration

Recall that we have to predict a value of p_n representing some information about a future outcome ω_n given past outcomes $\omega_1, \omega_2, \ldots, \omega_{n-1}$. In binary case, the number p_n can be interpreted as a probability of the event $\omega_n = 1$. It is easy to see that in this case the number p_n is also the mathematical expectation of a random variable ω_n taking values: 1 – with probability p_n, and 0 – with probability $1 - p_n$.

In the case of finite outcomes set, p_n can be a vector of probabilities of all possible outcomes.

In the measure-theoretic framework, we can suppose that p_n is a conditional probability distribution of a future outcome $\omega_n = 1$ given past outcomes $\omega_1, \omega_2, \ldots, \omega_{n-1}$.

But in reality, we should recognize that we have only individual sequence $\omega_1, \omega_2, \ldots, \omega_{n-1}$ of outcomes and that the corresponding forecasts p_n whose testing is considered may fall short of defining the full probability distribution in the whole space of infinite sequences of outcomes. In this section, we do not suppose that any such overall probability distribution exists.

At the same time, without a probabilistic model it is not obvious how to measure the performance of the method of prediction. We consider the case where there is no hypothesis about a mechanism generating outcomes ω_i. In this case we use different cost functionals and tests free from probability distribution to evaluate the performance of our forecasts.

A minimal requirement for testing any prediction algorithm is that it should be calibrated (see Dawid [11]). Dawid gave an informal explanation of calibration for binary outcomes. Let a sequence $\omega_1, \omega_2, \ldots, \omega_{n-1}$ of binary outcomes be observed by a forecaster whose task is to give a probability p_n of a future event $\omega_n = 1$. In a typical example, p_n is interpreted as a probability that it will rain. A forecaster is said to be well-calibrated if it rains as often as he leads us to expect. It should rain about 80% of the days for which $p_n = 0.8$, and so on.

The average deviation of the empirical frequency of the event $\omega_n = 1$ from the average value of predictions p_n such that $p_n \approx p^*$ for different values of p^* can be used as a test for rejecting "bad"

predictors.

The checking rule of weather forecaster can be written as follows: for any real number $p^* \in [0,1]$, it holds

$$\frac{\sum_{i=1\&p_i\approx p^*}^{n} \omega_i}{\sum_{i=1\&p_i\approx p^*}^{n} 1_{p_i\approx p^*}} \approx p^* \tag{3.1}$$

as the denominator of the (3.1) tends to infinity for $n \to \infty$. Here we used the symbol $\approx$ of approximate equality because in applications the number p^* may be specified only with some degree of accuracy. The condition $p_i \approx p^*$ requires further clarification.

In this section we consider a more general set of outcomes: now $\omega_n \in [0,1]$ for all n and the forecast $p_n \in [0,1]$ is interpreted as a mean value of a future outcome ω_n with respect to an unknown to us probability distribution. We do not know precise form of such distributions – we should predict only future means.

Consider the scheme of actions of *Forecaster* and *Nature* in the form of the following perfect-information protocol of a game between these two players.

FOR $n = 1, 2, \ \ldots$

Forecaster announces a forecast $p_n \in [0,1]$.

Nature announces an outcome $\omega_n \in \{0,1\}$.

ENDFOR

Since we consider a perfect information protocol (a game), *Forecaster* and *Nature* can use all the information known by the time of their action.

In particular, at step n, *Nature* can use the forecast p_n issued by *Forecaster*; *Forecaster* does not know the outcome ω_n, since at the moment of issue of the forecast p_n *Nature* has not yet announced its outcome.

Let us give a precise definition of calibration proposed by Dawid [12]. Consider any subintervals $I = [a,b]$, $(a,b]$, $[a,b)$, (a,b) of the unit interval $[0,1]$ and their characteristic functions

$$I(p) = \begin{cases} 1 \text{ if } p \in I, \\ 0 \text{ otherwise.} \end{cases}$$

A sequence of forecasts $p_1, p_2, \ldots$ is *well-calibrated* on an infinite sequence of outcomes $\omega_1, \omega_2, \ldots$ if for characteristic function $I(p)$ of

any subinterval $[0,1]$ *the calibration error* tends to zero, ie,

$$\frac{\sum_{i=1}^{n} I(p_i)(\omega_i - p_i)}{\sum_{i=1}^{n} I(p_i)} \longrightarrow 0 \tag{3.2}$$

as the denominator of the ratio (3.2) tends to zero for $n \to \infty$. Any characteristic function $I(p_i)$ defines some checking rule which selects time moments i, where we calculate the difference between the outcome ω_i and the forecast p_i.

A simple observation shows that any deterministic forecasting algorithm f will not always be calibrated. In particular, for any such forecasting algorithm we can define a sequence $\omega = \omega_1, \omega_2, \ldots$ such that

$$\omega_i = \begin{cases} 1 \text{ if } p_i < \frac{1}{2}, \\ 0 \text{ otherwise,} \end{cases}$$

where $p_i = f(\omega_1, \ldots, \omega_{i-1})$ are predictions computing by this algorithm, $i = 1, 2, \ldots$. Easy to see that for the interval $I = [0, \frac{1}{2})$ or for the interval $I = [\frac{1}{2}, 1)$, the condition of calibration (3.2) is violated.

The sequence $\omega = \omega_1, \omega_2, \ldots$ defined above is the simplest example of "an adversatively adaptive" strategy of *Nature*.

Generating the next outcome ω_i *Nature* already knows the forecast p_i and uses this knowledge to create the next outcome.

This example shows that "a universal" deterministic forecasting procedure does not exist. Such drawback can be overcome with the help of the notion of randomized forecasting system. Let $\mathcal{P}[0,1]$ be the set of all probability measures in the set $[0,1]$.

A randomized forecasting system is a function $f : \Xi \to \mathcal{P}[0,1]$ whose values are probability distributions in the unit interval $[0,1]$. Denote any such probability distribution $\Pr_x(\cdot) = f(x)$, where x is a finite sequence of outcomes.

Nature is *oblivious* if it does not use predictions made by *Forecaster*. In other words, *Nature* generate all outcomes in advance before the process of forecasting and reveals them step-by-step according to the protocol of the game.

We denote $\omega^{i-1} = \omega_1, \ldots, \omega_{i-1}$. In case of oblivious *Nature*, for any infinite sequence of outcomes $\omega = \omega_1, \omega_2, \ldots$, the independent conditional probabilities $\Pr_{\omega^{i-1}}(\cdot)$, $i = 1, 2, \ldots$, define an overall

probability distribution $\Pr = \prod_{i=1}^{\infty} \Pr_{\omega^{i-1}}$ in the set of all infinite trajectories of forecasts $p_1, p_2, \ldots$, where $p_i \in [0,1]$, $i = 1, 2, \ldots$. A sequence of outcomes ω is a parameter of this distribution.

An overall probability distribution Pr exists in a more general case of non-oblivious *Nature*. In this case the sequence of outcomes $\omega = \omega_1, \omega_2, \ldots$ issued by *Nature* depends on the sequence of predictions $p_1, p_2, \ldots$ made by *Forecaster*. More precise, any finite sequence of outcomes $\omega^n = \omega_1, \ldots, \omega_n$ is a measurable function of a sequence of predictions $p_1, \ldots, p_n$ for $n = 1, \ldots, n$. In this case by Ionesco–Tulcea theorem (see Schiryev [29]) an overall probability distribution exists such that

$$\Pr\{p_n \in A | p_1, \ldots, p_{n-1}\} = \Pr_{\omega^{n-1}}(A)$$

for any Borel set $A \subseteq [0,1]$ and for all trajectories of forecasts $p_1, p_2, \ldots$.

In all these cases we can consider the probability Pr of the event (3.2).

Foster and Fohra [13] and Kakade and Foster [17], defined a universal method of forecasting: given $\Delta > 0$ a randomizing forecasting system f can be constructed such that for any infinite bynary sequence $\omega = \omega_1, \omega_2, \ldots$:

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) \right| \leqslant \Delta,$$

with Pr-probability one, where trajectories of forecasts $\tilde{p}_1, \tilde{p}_2, \ldots$ are distributed by the probability distribution Pr and $I(p)$ is the characteristic function of any subinterval of the unit interval $[0,1]$.

3.3. Computing the well-calibrated forecasts

We present a modified version of the randomized forecasting algorithm of Kakade and Foster [17].

Unlike the previous section we consider real outcomes. Let $\omega_1, \omega_2, \ldots$ be an infinite sequence of elements of $[0,1]$ given online.

We construct an algorithm for computing the random forecasts $p_n \in [0,1]$ of future outcomes ω_n given the past outcomes $\omega_1, \ldots, \omega_{n-1}$. The main requirement for such forecasts: they should be well-calibrated with probability one. The corresponding probability distribution is the internal distribution constructed in the process of adaptation.

Define a partition of the unit interval $[0,1]$ on subintervals of length $\Delta = 1/K$ by means of rational points $v_i = i\Delta$, where $i = 0, 1, \ldots, K$. Let V be the set of all this points. Any number $p \in [0,1]$ can be represented as a linear combination of two boundary points of the subinterval containing p:

$$p = \sum_{v \in V} w_v(p) v = w_{v_{i-1}}(p) v_{i-1} + w_{v_i}(p) v_i,$$

where $p \in [v_{i-1}, v_i]$, $i = \lfloor p/\Delta + 1 \rfloor$, and

$$w_{v_{i-1}}(p) = 1 - \frac{p - v_{i-1}}{\Delta}, \; w_{v_i}(p) = 1 - \frac{v_i - p}{\Delta}.$$

Define $w_v(p) = 0$ for all other values $v \in V$.

In what follows a deterministic forecast p issued by the algorithm described below will be rounded to v_{i-1} with probability $w_{v_{i-1}}(p)$ and to v_i with probability $w_{v_i}(p)$.

We first construct an algorithm computing deterministic forecasts.

Suppose that forecasts $p_1, \ldots, p_{n-1}$ be already defined (let $p_1 = 0$).

Let us compute the forecast p_n. Define an auxiliary vector

$$\bar{\mu}_{n-1} = (\mu_{n-1}(v_0), \ldots, \mu_{n-1}(v_K)),$$

where

$$\mu_{n-1}(v) = \sum_{i=1}^{n-1} w_v(p_i)(\omega_i - p_i)$$

for $v \in V$. It holds

$$(\mu_n(v))^2 = (\mu_{n-1}(v))^2 + 2 w_v(p_n) \mu_{n-1}(v)(\omega_n - p_n) + \\ + (w_v(p_n))^2 (\omega_n - p_n)^2. \qquad (3.3)$$

Summing (3.3) over v, we obtain:

$$\sum_{v\in V}(\mu_n(v))^2 = \sum_{v\in V}(\mu_{n-1}(v))^2 + \\ +2(\omega_n - p_n)\sum_{v\in V} w_v(p_n)\mu_{n-1}(v) + \\ + \sum_{v\in V}(w_v(p_n))^2(\omega_n - p_n)^2. \tag{3.4}$$

Change the order of summation in the sum of auxiliary variables:

$$\sum_{v\in V} w_v(p)\mu_{n-1}(v) = \\ = \sum_{v\in V} w_v(p)\sum_{i=1}^{n-1} w_v(p_i)(\omega_i - p_i) = \\ = \sum_{i=1}^{n-1}(\sum_{v\in V} w_v(p)w_v(p_i))(\omega_i - p_i) = \\ = \sum_{i=1}^{n-1}(\bar{w}(p)\cdot\bar{w}(p_i))(\omega_i - p_i) = \\ = \sum_{i=1}^{n-1} K(p,p_i)(\omega_i - p_i),$$

where

$$\bar{w}(p) = (w_0,\ \ldots,\ w_{v_K}) = (0,\ \ldots,\ w_{v_{i-1}}(p), w_{v_i}(p),\ \ldots,\ 0)$$

be the vector of probabilities of rounding, $p \in [v_{i-1}, v_i]$, and

$$K(p,p_i) = (\bar{w}(p)\cdot\bar{w}(p_i)) \tag{3.5}$$

be the dot product of the corresponding vectors (a kernel). By definition $K(p,p_i)$ is a continuous function.

The second term on the right-hand side of the equality (3.4) can be made less than or equal to zero for an appropriate value of p_n. Indeed, we can define $p_n = 0$ if this term is negative for all $p \in [0,1]$

and $p_n = 1$ if this term is positive for all $p \in [0,1]$. Otherwise, define p_n be equal to some root $p_n = p$ of the equation:

$$\sum_{v \in V} w_v(p)\mu_{n-1}(v) = \sum_{i=1}^{n-1} K(p, p_i)(\omega_i - p_i) = 0. \tag{3.6}$$

Such a root exists by the intermediate value theorem. We call p_n the deterministic forecast.

The third term of (3.4) is bounded by 1. Indeed, since $|\omega_i - p_i| \leqslant 1$ for all i, we have for any n:

$$\sum_{v \in V} (w_v(p_n))^2(\omega_n - p_n)^2 \leqslant \sum_{v \in V} w_v(p_n) = 1.$$

Therefore, the forecasts p_i satisfy:

$$\sum_{v \in V} (\mu_n(v))^2 \leqslant \sum_{i=1}^{n} \sum_{v \in V} (w_v(p_i))^2(\omega_i - p_i)^2 \leqslant n.$$

Let now $\tilde{p}_i$ be a random variable taking values $v \in V$ with probabilities $w_v(p_i)$. [1] Let also, $I(p)$ be the characteristic function of any subinterval of $[0,1]$. For any i, the mathematical expectation of the random variable $I(\tilde{p}_i)(\omega_i - \tilde{p}_i)$ is equal:

$$E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i)) = \sum_{v \in V} w_v(p_i) I(v)(\omega_i - v). \tag{3.7}$$

By the strong martingale law of large numbers (see Corollary 8.7 below), with Pr-probability one:

$$\left| \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) - \frac{1}{n} \sum_{i=1}^{n} E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i)) \right| \to 0 \tag{3.8}$$

as $n \to \infty$.

By definition of deterministic forecast p_i and $w_v(p)$:

$$\left| \sum_{v \in V} w_v(p_i) I(v)(\omega_i - v) - \sum_{v \in V} w_v(p_i) I(v)(\omega_i - p_i) \right| < \Delta \tag{3.9}$$

[1] Only $w_{v_i}(p)$ and $w_{v_{i+1}}(p)$ are nonzero, where $p_n \in [v_i, v_{i+1}]$.

for each i.

Applying the Cauchy–Schwarz inequality to vectors $\bar{\mu}_n = \{\mu_n(v) : v \in V\}$ and $\{I(v) : v \in V\}$ and taking into account (3.9), we obtain:

$$\begin{aligned}\left|\sum_{i=1}^{n}\sum_{v\in V} w_v(p_i)I(v)(\omega_i - p_i)\right| = \\ = \left|\sum_{v\in V} I(v)\sum_{i=1}^{n} w_v(p_i)(\omega_i - p_i)\right| \leqslant \\ \leqslant \sqrt{\sum_{v\in V}(\mu_n(v))^2}\sqrt{\sum_{v\in V} I(v)} \leqslant \\ \leqslant \sqrt{(K+1)n}, \qquad (3.10)\end{aligned}$$

where $K = 1/\Delta$ is the cardinality of the partition.

Using (3.9) and (3.10), we obtain the upper bound:

$$\begin{aligned}\left|\sum_{i=1}^{n} E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i))\right| = \\ = \left|\sum_{i=1}^{n}\sum_{v\in V} w_v(p_i)I(v)(\omega_i - v)\right| \leqslant \qquad (3.11) \\ \leqslant \Delta n + \sqrt{n(1+1/\Delta)} \qquad (3.12)\end{aligned}$$

for all n.

By (3.12) and (3.8) we obtain that, with Pr-probability one:

$$\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i)\right| \leqslant \Delta. \qquad (3.13)$$

We formulate the main result of this section in the following theorem.

Theorem 3.1. *For any $\Delta > 0$, a randomized forecasting system f can be constructed such that for any infinite sequence of outcomes $\omega = \omega_1, \omega_2, \ldots$, with* Pr*-probability one:*

$$\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i)\right| \leqslant \Delta,$$

where infinite trajectories of forecasts $\tilde{p}_1, \tilde{p}_2, \ldots$ *are distributed by the corresponding to* f *overall probability distribution* Pr*,* $I(p)$ *is the characteristic function of any subinterval of* $[0,1]$*. We call such a function the checking rule.*

Using variable precision of rounding $\Delta = \Delta_s$, where $\Delta_s \to 0$ as $s \to \infty$, we can obtain asymptotic result:

Theorem 3.2. *A randomized forecasting system can be constructed such that for any infinite sequence* $\omega = \omega_1, \omega_2, \ldots,$ *with* Pr*-probability one,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) = 0,$$

where $I(p)$ *is the characteristic function of any subinterval of* $[0,1]$*.*

The proof of this theorem is similar to the proof of Theorem 3.5 of Section 3.5.

3.4. Defensive forecasting

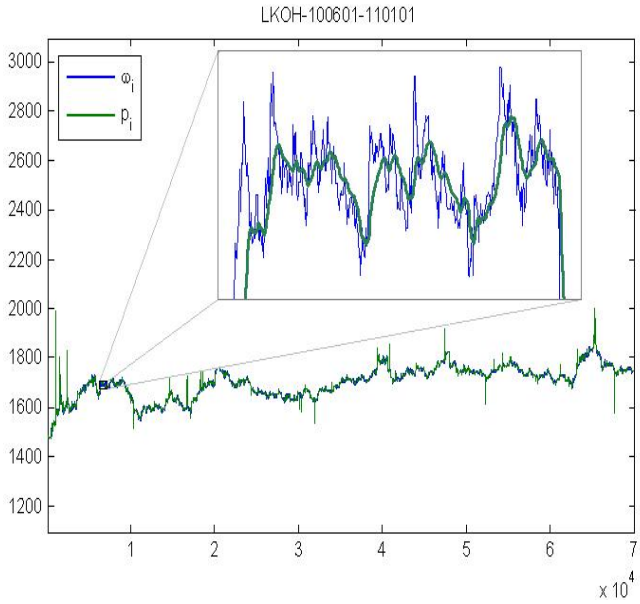
We consider two approaches to universal prediction:

– universal prediction, in which a probability distribution in the set of all possible predictions is issued as a forecast; in this case arbitrary subintervals I of the unit interval are used as checking rules $I(p)$;

– universal prediction, in which forecasts are deterministic but continuous functions (in particular, continuous approximations of characteristic functions of subintervals) serve as analogs of checking rules.

In the first case, for any sequences of outcomes given online we have constructed in the previous section a sequence of forecasts satisfying the condition of calibration with probability one.

In the second case, we will construct in this section a sequence of deterministic forecasts satisfying the condition of calibration, were characteristic functions of subintervals are replaced by continuous weights.



. 2.1. Example of a sequence of outcomes $\omega_1, \omega_2, \ \ldots$ and well-calibrated forecasts $p_1, p_2, \ldots$

In some sense both approaches are equivalent (see Kakade and Foster [17]).

In this section we consider the second approach. Vovk in [41] and [44] generalized the method of universal forecasting discovered by Foster and Vohra [13] and Kakade and Foster [17] to arbitrary RKHS. We present an idea of this generalization.

We formulate the problem of prediction as a game between players: *Nature*, *Forecaster* and *Skeptic*.

In this game, forecasts will be deterministic like forecasts p_i computed as roots of equations (3.6) in Section 3.3. We consider a more general setting, namely, we add a side information named *signals*.

Let $X \subseteq \mathcal{R}^m$ be the set of all m-dimensional vectors $\bar{x} =$

$(x_1, \ \ldots, \ x_m)$ and

$$\|\bar{x}\| = \sqrt{\sum_{i=1}^{m} x_i^2}$$

be the Euclidian norm. We call these vectors *signals*.

Define $\mathcal{K}_0 = 1$.

The game is regulated by the following perfect-information protocol:
FOR $n = 1, 2, \ \ldots$
Nature announces a signal $\bar{x}_n \in X$.
Skeptic announces a continuous function $S_n : [0, 1] \to \mathcal{R}$.
Forecaster announces a forecast $p_n \in [0, 1]$.
Nature announces an outcome $y_n \in \{0, 1\}$.
Skeptic updates his capital:
$\mathcal{K}_n = \mathcal{K}_{n-1} + S_n(p_n)(y_n - p_n)$.
ENDFOR

The next theorem shows that *Forecaster* has a strategy such that *Skeptic's* gain non-increases in the process of the game.

Theorem 3.3. *(Vovk et al. [42]) Forecaster has a strategy such that* $\mathcal{K}_0 \geqslant \mathcal{K}_1 \geqslant \ldots \mathcal{K}_n \geqslant \ldots$

Proof. *Forecaster* computes a forecast p_n at any step n as follows.

If $S_n(p) > 0$ for all $p \in [0, 1]$ then define $p_n = 1$. If $S_n(p) < 0$ for all $p \in [0, 1]$ then define $p_n = 0$. Otherwise, by the intermediate value theorem equation

$$S_n(p) = 0, \tag{3.14}$$

has a root p; in this case let p_n be some root of (3.14).

Evidently, in this case the *Skeptic's* gain non-increases for any choice of continuous function $S_n(p)$, ie,

$$\mathcal{K}_0 \geqslant \mathcal{K}_1 \geqslant \ldots \mathcal{K}_n \geqslant \ldots$$

holds for all n. $\triangle$

We use a kernel $K((p,\bar{x}),(p',\bar{x}'))$ which is a continuous real function on $([0,1]\times X)^2$. Example of a kernel is the Gaussian kernel:

$$K((p,\bar{x}),(p',\bar{x}')) = \\ = \exp\left(-\frac{(p-p_i)^2}{\sigma_1^2} - \frac{\|\bar{x}-\bar{x}'\|^2}{\sigma_2^2}\right), \tag{3.15}$$

where σ_1, σ_2 are parameters.

Consider a *Skeptic's* strategy which forces *Forecaster* to make "well calibrated" forecasts at any step n independently of how *Nature* outputs her outcomes. Let the forecasts $p_1,\dots,p_{n-1}$ be given at the beginning of step n. Define the function

$$S_n(p) = \sum_{i=1}^{n-1} K((p,\bar{x}_n),(p_i,\bar{x}_i))(y_i-p_i).$$

Let *Forecaster* uses the strategy defined in Theorem 3.3. Then *Skeptic's* gain for N steps satisfies:

$$\mathcal{K}_N - \mathcal{K}_0 = \sum_{n=1}^{N} S_n(p_n)(y_n-p_n) = \\ = \sum_{n=1}^{N}\sum_{i=1}^{n-1} K((p_n,\bar{x}_n),(p_i,\bar{x}_i))(y_i-p_i)(y_n-p_n) = \\ = \frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{N} K((p_n,\bar{x}_n),(p_i,\bar{x}_i))(y_i-p_i)(y_n-p_n) - \\ -\frac{1}{2}\sum_{n=1}^{N} K((p_n,\bar{x}_n),(p_n,\bar{x}_n))(y_n-p_n)^2. \tag{3.16}$$

By the theory presented in Section 2.5 a Hilbert feature space H and a mapping $\Phi : [0,1]\times X \to \mathcal{H}$ exist such that

$$K(a,b) = (\bar{\Phi}(a)\cdot\bar{\Phi}(b))$$

for $a,b\in[0,1]\times X$, where "$\cdot$" is the dot product in the space $\mathcal{H}$ (in what follows $\|\cdot\|_{\mathcal{H}}$ is the corresponding norm).

The quantity $c_{\mathcal{H}} = \sup\limits_{a} \|\Phi(a)\|_{\mathcal{H}}$ is called the embedding constant. Assume that the Hilbert space H has a finite embedding constant: $c_{\mathcal{H}} < \infty$. Rewrite (3.16) in the form:

$$\mathcal{K}_N - \mathcal{K}_0 = \frac{1}{2}\left\|\sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n)\right\|_{\mathcal{H}}^2 - \\ -\frac{1}{2}\sum_{n=1}^{N} \|\bar{\Phi}((p_n, \bar{x}_n)(y_n - p_n)\|_{\mathcal{H}}^2. \tag{3.17}$$

By the assumption

$$c_{\mathcal{H}} = \sup_{p,\bar{x}} \|\bar{\Phi}(p, \bar{x})\|_{\mathcal{H}} < \infty.$$

By Theorem 3.3 the inequality $\mathcal{K}_N - \mathcal{K}_0 \leqslant 0$ holds for all n. Then by (3.17):

$$\frac{1}{2}\left\|\sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n)\right\|_{\mathcal{H}}^2 \leqslant \frac{1}{2}NC^2. \tag{3.18}$$

Rewrite the inequality (3.18) in the form:

$$\left\|\sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n)\right\|_{\mathcal{H}} \leqslant \sqrt{N}C. \tag{3.19}$$

In other words, the mean error of the forecasting algorithm is bounded:

$$\frac{1}{N}\left\|\sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n)\right\|_{\mathcal{H}} \leqslant \frac{C}{\sqrt{N}}.$$

Using this bound, we can obtain a condition of calibration similar to the condition from Section 3.3. To do this, consider some family of smooth approximation for the characteristic functions of singletons $\{(p^*, \bar{x}^*)\}$, where $(p^*, \bar{x}^*) \in [0, 1] \times X$, ie, functions of the form:

$$K((p^*, \bar{x}^*), (p, \bar{x})) = I_{(p^*, \bar{x}^*)}(p, \bar{x}). \tag{3.20}$$

We assume that this function is a kernel. An example of such family $I_{p^*}(p)$ is the family of Gaussian kernels (3.15).

The forecasts p_i satisfy the following inequality:

Corollary 3.1.

$$\left|\frac{1}{N}\sum_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)(y_n-p_n)\right| \leqslant \frac{C^2}{\sqrt{N}} \tag{3.21}$$

for each point $(p^*,\bar{x}^*)\in[0,1]\times X$.

Proof. For any kernel a function $\Phi(p,\bar{x})$ exists which has a range in a Hilbert feature space H such that

$$K((p^*,\bar{x}^*),(p,\bar{x}))=I_{(p^*,\bar{x}^*)}(p,\bar{x})=(\bar{\Phi}(p^*,\bar{x}^*)\cdot\bar{\Phi}(p,\bar{x})).$$

Applying Cauchy–Schwarz inequality for (3.19), we obtain:

$$\left|\sum_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)(y_n-p_n)\right| =$$

$$=\left|\left(\left(\sum_{n=1}^{N}\bar{\Phi}(p_n,\bar{x}_n)(y_n-p_n)\right)\cdot\bar{\Phi}(p^*,\bar{x}^*)\right)\right|\leqslant$$

$$\leqslant\left\|\sum_{n=1}^{N}\bar{\Phi}(p_n,\bar{x}_n)(y_n-p_n)\right\|_{\mathcal{H}}\|\bar{\Phi}(p^*,\bar{x}^*)\|_{\mathcal{H}}\leqslant C^2\sqrt{N}.$$

From this we obtain (3.21). $\triangle$

The quantity:

$$\sum_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)$$

is a smooth analogue of the total number of pairs (p_n,x_n) locating in in the "soft" neighborhood of the pair $(p^*,\bar{x}^*)$.

The inequality (3.21) can be rewritten in the form:

$$\left|\frac{\sum\limits_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)(y_n-p_n)}{\sum\limits_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)}\right|\leqslant\frac{C^2\sqrt{N}}{\sum\limits_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)}. \tag{3.22}$$

The bound (3.22) is valid for

$$\sum_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n, \bar{x}_n) \gg \sqrt{N},$$

ie, the convergence of frequencies to predictions take place only for subsequences of "statistically significant" length.

The universal forecasting algorithm presented in this section can easily implemented as a computer program.

3.5. Universal algorithmic trading

In this section, we consider financial applications of the method of calibration. [2] The method of universal forecasting will be applied to the problem of universal sequential investment in Stock Market. We consider the method of trading called in financial industrial applications *algorithmic trading* or *systematic quantitative trading*, which means rule-based automatic trading strategies, usually implemented with computer based trading systems.

A non-traditional objective (in computational finance) is to develop algorithmic trading strategies that are, in some sense, always guaranteed to perform well. In competitive analysis, the performance of an algorithm is measured to any trading algorithm from a broad class. We only ask than an algorithm performs well, relative to the difficulty in classifying of the input data. Given a particular performance measure, an adaptive algorithm is strongly competitive with a class of trading algorithms if it achieves the maximum possible regret over all input sequences. Unlike the statistical theory, no stochastic assumptions are made about the stock prices.

This line of research in finance was pioneered by Cover [9] who designed universal portfolio selection algorithms that are proven to perform well (in terms of their total return) with respect to some adaptive online or offline benchmark algorithms. Such algorithms are called *universal*. We present this algorithm in Section 5.8 below.

[2] This section is technical and can be passed on the first reading.

In this framework, we consider a universal trading for one stock. We construct "a universal" strategy for algorithmic trading in Stock Market which is performed at least as well as any trading strategy that is not excessively complex. By "performance" we mean return per unit of currency on an investment.

Trading in Stock Market.

The process of trading proceeds as follows: observing a sequence of past prices of a stock and the side information, a forecaster assigns an estimate to a future price. Then, using this forecast, a trader makes a decision to buy or sell shares of the stock. He chooses a strategy: going long or going short, or skip the step. In finance, a long position in a security, such as a stock or a bond, or equivalently to be long in a security, means the holder of the position owns the security and will profit if the price of the security goes up. Short selling (also known as shorting or going short) is the practice of selling securities or other financial instruments, with the intention of subsequently repurchasing them ("covering") at a lower price.

The forecasting method is defined as a combination of the randomized forecasting defined in Section 3.3 and defensive forecasting defined in Section 3.4. We use also Dawid's notion of calibration considered in Section 3.2 with more general checking rules.

Theorem 3.4, says that this trading strategy is universal – it performs asymptotically at least as well as any trading strategy presented by any continuous function from a piece of side information. This method and the corresponding numerical experiments are presented in V'yugin and Trunov [45].

Suppose that real numbers $S_1, S_2, \ldots,$ that are interpreted as prices of a stock, are given online. We assume that they are bounded and scaled such that $0 \leqslant S_i \leqslant 1$ for all i. We present the process of trading in a stock market in the form of the protocol of a game with two players called *traders*. *Trader M* uses the randomized strategy that is a random variable $\tilde{M}_i$. *Trader D* uses an arbitrary stationary trading strategy D that is a real function defined on $[0,1]$.

In general, under the strategy we mean an algorithm (possibly randomized), which at each step i of the game outputs the number of units of the financial instrument you want to buy (if the number

is positive or equal to zero) or sell (if it is negative). [3] For *Trader M*, this number is a value of the random variable $\tilde{M}_i$ and for *Trader D*, this number equals $D(x_i)$.

At the beginning of each step i, traders are given some data x_i that is relevant to predicting a future price S_i of the stock. We call x_i *a signal* or a piece of *side information*. The real number x_i belongs to $[0, 1]$ and can encode any information. For example, it can even be the future price S_i.

Trader D uses at any step i only information x_i – he buys (or sells) $D(x_i)$ units of shares. The strategy of this type will be called *stationary*.

For *Trader M*, this game is a game with perfect information. *Trader M*, for defining the random variable $\tilde{M}_i$, may use all values of S_j, x_j for $j \leqslant i$, as well as their randomized values.

Past stock prices, signals, and predictions can be encoded in the signal x_i, so the *Trader D* can also use this information. There is a restriction – a method of encoding must be defined in advance at the beginning of the game. Another limitation for *Trader D* – the function D must be continuous.

Method of randomization revisited.

We will use the method of randomization defined in Section 3.3. Some special kernel corresponds to the method of randomization defined below.

A random variable $\tilde{y}$ is called randomization of a real number $y \in [0, 1]$ if $E(\tilde{y}) = y$, where E is the symbol of mathematical expectation with respect to the probability distribution corresponding to $\tilde{y}$.

We use a specific method of randomization of real numbers from the unit interval defined in Section 3.3. Given positive integer number K divide the interval $[0, 1]$ on subintervals of length $\Delta = 1/K$ with rational endpoints $v_i = i\Delta$, where $i = 0, 1, \ldots, K$. Let V denotes the set of these points. Any number $p \in [0, 1]$ can be represented as a linear combination of two neighboring endpoints of V defining

[3] We believe that the number of units of a financial instrument purchased by traders may take any real value.

subinterval containing p :

$$p = \sum_{v \in V} w_v(p)v = w_{v_{i-1}}(p)v_{i-1} + w_{v_i}(p)v_i, \tag{3.23}$$

where $p \in [v_{i-1}, v_i]$, $i = \lfloor p^1/\Delta + 1 \rfloor$, $w_{v_{i-1}}(p) = 1 - (p - v_{i-1})/\Delta$, and $w_{v_i}(p) = 1 - (v_i - p)/\Delta$. Define $w_v(p) = 0$ for all other $v \in V$. Define a random variable:

$$\tilde{p} = \begin{cases} v_{i-1} \text{ with probability } w_{v_{i-1}}(p) \\ v_i \text{ with probability } w_{v_i}(p) \end{cases}$$

Let $\bar{w}(p) = (w_v(p) : v \in V)$ be a vector of probabilities of rounding. For $z, z' \in [0, 1]$, define the dot product $K(z, z') = (\bar{w}(z) \cdot \bar{w}(z'))$ which is a kernel function. [4]

In what follows we will consider a variable parameter Δ. More correctly, we will define a sequence of parameters $\Delta_1 \geqslant \Delta_2 \geqslant \cdots \to 0$. At each step i of the construction, we will round off real numbers up to Δ_i. This method of random rounding will be called *sequential randomization*.

Universal trading strategy.

Define a universal trading strategy as a sequence of random variables $\tilde{M}_i$, $i = 1, 2, \ldots$ as follows. Using algorithm presented in Section 3.5.1, at each step i, we compute a forecast p_i of a future price S_i and randomize it to $\tilde{p}_i$. We also randomize the past price S_{i-1} of the stock to $\tilde{S}_{i-1}$. Define the random variable:

$$\tilde{M}_i = \begin{cases} 1 \text{ if } \tilde{p}_i > \tilde{S}_{i-1}, \\ -1 \text{ otherwise.} \end{cases}$$

Trading game.

In case $\tilde{M}_i > 0$ *Trader M* going long, and going short, otherwise. The same holds for *Trader D*. We suppose that traders can borrow money for buying shares and can incur debt. The core of universal strategy is the algorithm for computing the well-calibrated forecasts $\tilde{p}_i$. This algorithm is presented in Section 3.5.1.

[4] Many other methods of randomization also work.

FOR $i = 1, 2 \ldots$
Stock Market announces a signal $x_i \in [0, 1]$.
Trader M bets by buying (or selling) the random number $\tilde{M}_i$ of shares of the stock by S_{i-1} each.
Trader D bets by buying (or selling) a number $D(x_i)$ of shares of the stock by S_{i-1} each.
Stock Market reveals a price S_i of the stock.
Trader M updates his total gain (loss):
$\mathcal{K}_i^M = \mathcal{K}_{i-1}^M + \tilde{M}_i(S_i - S_{i-1})$. We get $\mathcal{K}_0^M = 0$.
Trader D updates his total gain (loss):
$\mathcal{K}_i^D = \mathcal{K}_{i-1}^D + D(x_i)(S_i - S_{i-1})$. We get $\mathcal{K}_0^D = 0$.
ENDFOR

Figure 3.1: Protocol of the trading game

Trader M can buy or sell only one share of the stock. Therefore, in order to compare the performance of the traders we have to standardize the strategy of *Trader D*. Recall the norm $\|D\|_\infty = \sup_{0 \leqslant x \leqslant 1} |D(x)|$, where D is a continuous function. We will use $\|D\|_\infty^{-1}$ as a normalization factor.

Assume that $S_1, S_2, \cdots \in [0, 1]$ and $x_1, x_2, \cdots \in [0, 1]$ be given sequentially according to the protocol presented on Fig 3.1.

Main result of this section is presented in the following theorem. It says that, with probability one, the average gain of the universal trading strategy is asymptotically not less than the average gain of any stationary trading strategy from one share of the stock:

Theorem 3.4. *An algorithm for computing forecasts and a sequential method of randomization can be constructed such that for any continuous nonzero function D:*

$$\liminf_{n \to \infty} \frac{1}{n} \left(\mathcal{K}_n^M - \|D\|_\infty^{-1} \mathcal{K}_n^D \right) \geqslant 0 \tag{3.24}$$

holds almost surely with respect to a probability distribution generated by the corresponding sequential randomization.

The proof of this theorem is given in Section 3.5.2, where we

construct the corresponding optimal trading strategy based on the well-calibrated forecasts defined in Section 3.5.1.

Reproducing Kernel Hilbert Spaces revisited.

First, we will prove in Section 3.5.2 that the trading strategy $\tilde{M}_i$ is universal with respect to all stationary trading strategies from any benchmark class called RKHS (Reproducing Kernel Hilbert Space). After that, using a universal RKHS (that will be defined below), we extend the universality property to the class of all continuous stationary trading strategies.

Recall that a Hilbert space $\mathcal{F}$ of real-valued functions on a compact set X is called RKHS on X if the evaluation functional $f \to f(x)$ is continuous for each $x \in X$. Let $\|\cdot\|_{\mathcal{F}}$ be a norm on $\mathcal{F}$ and $c_{\mathcal{F}}(x) = \sup_{\|f\|_{\mathcal{F}} \leqslant 1} |f(x)|$. The embedding constant of $\mathcal{F}$ is defined: $c_{\mathcal{F}} = \sup_x c_{\mathcal{F}}(x)$.

We consider RKHS $\mathcal{F}$ on $X = [0,1]$ with $c_{\mathcal{F}} < \infty$. An example of such RKHS is the Sobolev space $\mathcal{F} = H^1([0,1])$, which consists of absolutely continuous functions $f : [0,1] \to \mathcal{R}$ with $\|f\|_{\mathcal{F}} < \infty$, where $\|f\|_{\mathcal{F}} = \sqrt{\int_0^1 (f(t))^2 dt + \int_0^1 (f'(t))^2 dt}$. For this space, $c_{\mathcal{F}} = \sqrt{\coth 1}$ (see [41]).

Let $\mathcal{F}$ be an RKHS on X with the dot product $(f \cdot g)$ for $f, g \in \mathcal{F}$. By Riesz–Fisher theorem, for each $x \in X$ there exists $k_x \in \mathcal{F}$ such that $f(x) = (k_x \cdot f)$. The reproduced kernel is defined $K(x,y) = (k_x \cdot k_y)$.

Conversely, any kernel $K(x,y)$ defines some *canonical* RKHS $\mathcal{F}$ and a mapping $\Phi : X \to \mathcal{F}$ such that $K(x,y) = (\Phi(x) \cdot \Phi(y))$.

3.5.1. Well-calibrated forecasting with side information

In this section, we define an algorithm for computing the well-calibrated forecasts, which is a core of the investment strategy $\tilde{M}_i$.

We consider checking rules of a more general type than those used in Sections 3.2 and 3.3. For any subset $S \subseteq [0,1] \times [0,1]$ define

$$I_S(p,x) = \begin{cases} 1, \text{ if } (p,x) \in S, \\ 0, \text{ otherwise,} \end{cases}$$

where $p, x \in [0,1]$. In Section 3.5.2 we get $S = \{(p,x) : p > x\}$.

The following theorem is the main tool for analysis presented in Section 3.5.2.

Let $\mathcal{F}$ be an RKHS on $[0,1]$ with a finite embedding constant $c_{\mathcal{F}}$, $\|\cdot\|_{\mathcal{F}}$ be the norm, and $M(x,x')$ be the kernel on $[0,1]$.

Suppose that $S_1, S_2, \dots \in [0,1]$ be a sequence of real numbers and $x_1, x_2, \dots \in [0,1]$ be a sequence of signals given sequentially according to the protocol presented on Fig 3.1.

Theorem 3.5. *Given $\epsilon > 0$, an algorithm for computing forecasts $p_1, p_2, \dots$ and a sequential method of randomization can be constructed such that two conditions hold:*

- *for any $\delta > 0$, for any $S \subseteq [0.1]^2$ and for any n, with Pr-probability at least $1 - \delta$:*

$$\left|\sum_{i=1}^{n} I_S(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)\right| \leqslant$$
$$\leqslant 18 n^{3/4+\epsilon}(c_{\mathcal{F}}^2 + 1)^{1/4} + \sqrt{\frac{n}{2}\ln\frac{2}{\delta}}, \tag{3.25}$$

  where $\tilde{p}_1, \tilde{p}_2, \dots$ are the corresponding randomizations of $p_1, p_2, \dots$ and $\tilde{z}_1, \tilde{z}_2, \dots$ are randomizations of reals $z_1, z_2, \dots$ and $z_i = S_{i-1}$, $i = 1, 2, \dots$;

- *for any $D \in \mathcal{F}$:*

$$\left|\sum_{i=1}^{n} D(x_i)(S_i - p_i)\right| \leqslant \|D\|_{\mathcal{F}}\sqrt{(c_{\mathcal{F}}^2 + 1)n} \tag{3.26}$$

  for all n.

Proof. At first, given $\Delta > 0$, we modify a randomized rounding algorithm from previous section to construct some forecasting algorithm calibrated up to a precision Δ, and combine it with defensive forecasting algorithm. After that, we apply to this algorithm some "doubling trick" argument such that (3.25) will hold.

Proposition 3.1. *Under the assumption of Theorem 3.5, an algorithm for computing forecasts and a method of randomization can be constructed such that the inequality (3.26) holds for all D from RKHS $\mathcal{F}$ and for all n. Also, for any $\delta > 0$, S and n, with probability at least $1 - \delta$:*

$$\left|\sum_{i=1}^{n} I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)\right| \leqslant$$
$$\Delta n + 2\sqrt{\frac{n(c_{\mathcal{F}}^2 + 1)}{\Delta}} + \sqrt{\frac{n}{2}\ln\frac{2}{\delta}}.$$

Proof. Assume that the deterministic forecasts $p_1, \ldots, p_{n-1}$ be already defined (put $p_1 = 1/2$). We define a deterministic forecast p_n and randomly round it to $\tilde{p}_n$.

The partition $V = \{v_0, \ldots, v_K\}$ and probabilities of rounding were defined above by (3.23). We round p_n to v_{i-1} with probability $w_{v_{i-1}}(p_n)$ and to v_i with probability $w_{v_i}(p_n)$. We also randomly round $z_n = S_{n-1}$ to v_{s-1} with probability $w_{v_{s-1}}(z_n)$ and to v_s with probability $w_{v_s}(z_n)$, where $z_n \in [v_{s-1}, v_s]$.

Let $W_v(p_n, z_n) = w_{v^1}(p_n) w_{v^2}(z_n)$, where $v = (v^1, v^2)$ and $v^1, v^2 \in V$, and $W(p_n, z_n) = (W_v(p_n, z_n) : v \in V^2)$ be a vector of probability distribution in V^2. Define the corresponding kernel $K(p, z, p', z') = (W(p, z) \cdot W(p', z'))$.

By definition, the kernel $M(x, x')$ can be represented as a dot product in some feature space: $M(x, x') = (\Phi(x) \cdot \Phi(x'))$. Consider the function

$$U_n(p) = \sum_{i=1}^{n-1} (K(p, z_n, p_i, z_i) + M(x_n, x_i))(S_i - p_i). \tag{3.27}$$

Recall Theorem 3.3 which gives a general method for computing deterministic forecasts:

A sequence of forecasts $p_1, p_2, \ldots$ can be computed such that $\mathcal{M}_n \leqslant \mathcal{M}_{n-1}$ for all n, where $\mathcal{M}_0 = 1$ and $\mathcal{M}_n = \mathcal{M}_{n-1} + U_n(p_n)(S_n - p_n)$ for all i.

Indeed, if $U_n(p) > 0$ for all $p \in [0, 1]$ then define $p_n = 1$; if $U_n(p) < 0$ for all $p \in [0, 1]$ then define $p_n = 0$. Otherwise, define p_n

to be some root of the equation $U_i(p) = 0$ (some root exists by the intermediate value theorem). Evidently, $\mathcal{M}_n \leqslant \mathcal{M}_{n-1}$ for all n.

Now we continue the proof of Proposition 3.1.

We have for any N:

$$
\begin{aligned}
0 \geqslant \mathcal{M}_N - \mathcal{M}_0 = \sum_{n=1}^{N} U_n(p_n)(S_n - p_n) = \\
= \sum_{n=1}^{N} \sum_{i=1}^{n-1} (K(p_n, z_n, p_i, z_i) + M(x_n, x_i)) \times \\
\times (S_i - p_i)(S_n - p_n) = \\
= \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} K(p_n, z_n, p_i, z_i)(S_i - p_i)(S_n - p_n) - \\
- \frac{1}{2} \sum_{n=1}^{N} (K(p_n, z_n, p_n, z_n)(S_n - p_n))^2 + \\
+ \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} M(x_n, x_i)(S_i - p_i)(S_n - p_n) - \\
- \frac{1}{2} \sum_{n=1}^{N} (M(x_n, x_n)(S_n - p_n))^2 = \qquad (3.28) \\
= \frac{1}{2} \left\| \sum_{n=1}^{N} W(p_n, z_n)(S_n - p_n) \right\|^2 - \\
- \frac{1}{2} \sum_{n=1}^{N} \|W(p_n, z_n)\|^2 (S_n - p_n)^2 + \qquad (3.29) \\
+ \frac{1}{2} \left\| \sum_{n=1}^{N} \Phi(x_n)(S_n - p_n) \right\|_{\mathcal{F}}^2 - \\
- \frac{1}{2} \sum_{n=1}^{N} \|\Phi(x_n)\|_{\mathcal{F}}^2 (S_n - p_n)^2. \qquad (3.30)
\end{aligned}
$$

In (3.29), $\|\cdot\|$ is the Euclidian norm, and in (3.30), $\|\cdot\|_{\mathcal{F}}$ is the norm in RKHS $\mathcal{F}$.

Since $(S_n - p_n)^2 \leqslant 1$ for all n and:

$$\|W(p_n, z_n)\|^2 = \sum_{v \in V^2} (W_v(p_n, z_n))^2 \leqslant$$
$$\leqslant \sum_{v \in V^2} W_v(p_n, z_n) = 1,$$

the subtracted sum of (3.29) is upper bounded by N.

Since $\|\Phi(x_n)\|_{\mathcal{F}} = c_{\mathcal{F}}(x_n)$ and $c_{\mathcal{F}}(x) \leqslant c_{\mathcal{F}}$ for all x, the subtracted sum of (3.30) is upper bounded by $c_{\mathcal{F}}^2 N$. As a result we obtain:

$$\left\| \sum_{n=1}^{N} W(p_n, z_n)(S_n - p_n) \right\| \leqslant \sqrt{(c_{\mathcal{F}}^2 + 1)N} \tag{3.31}$$

$$\left\| \sum_{n=1}^{N} \Phi(x_n)(S_n - p_n) \right\|_{\mathcal{F}} \leqslant \sqrt{(c_{\mathcal{F}}^2 + 1)N} \tag{3.32}$$

for all N. Let us denote: $\bar{\mu}_n = \sum\limits_{i=1}^{n} W(p_i, z_i)(S_i - p_i)$. By (3.31), $\|\bar{\mu}_n\| \leqslant \sqrt{(c_{\mathcal{F}}^2 + 1)n}$ for all n.

Let $\bar{\mu}_n = (\mu_n(v) : v \in V^2\}$. By definition for any v:

$$\mu_n(v) = \sum_{i=1}^{n} W_v(p_i, z_i)(S_i - p_i). \tag{3.33}$$

Insert the term $I(v)$ in the sum (3.33), where I is the characteristic function of an arbitrary set $\mathcal{S} \subseteq [0,1] \times [0,1]$, sum by $v \in V^2$, and exchange the order of summation. Using Cauchy–Schwarz inequality for vectors $\bar{I} = (I(v) : v \in V^2)$, $\bar{\mu}_n = (\mu_n(v) : v \in V^2)$ and Euclidian norm, we obtain:

$$\left| \sum_{i=1}^{n} \sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - p_i) \right| =$$
$$= \left| \sum_{v \in V^2} I(v) \sum_{i=1}^{n} W_v(p_i, z_i)(S_i - p_i) \right| =$$
$$= (\bar{I} \cdot \bar{\mu}_n) \leqslant \|\bar{I}\| \cdot \|\bar{\mu}_n\| \leqslant \sqrt{|V^2|(c_{\mathcal{F}}^2 + 1)n} \tag{3.34}$$

for all n, where $|V^2| = (1/\Delta + 1)^2 \leqslant 4/\Delta^2$ is the cardinality of the partition.

Let $\tilde{p}_i$ be a random variable taking values $v \in V$ with probabilities $w_v(p_i)$. Recall that $\tilde{z}_i$ is a random variable taking values $v \in V$ with probabilities $w_v(z_i)$.

Let $\mathcal{S} \subseteq [0,1] \times [0,1]$ and I be its indicator function. For any i the mathematical expectation of a random variable $I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)$ is equal to:

$$\begin{aligned} &E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)) = \\ &= \sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - v^1), \end{aligned} \tag{3.35}$$

where $v = (v^1, v^2)$. Using Corollary 8.5 from the Azuma–Hoeffding inequality (see Lemma 8.2 of Section 8.6 below), we obtain for any $\delta > 0$, S and n, with Pr-probability $1 - \delta$:

$$\begin{aligned} \left| \sum_{i=1}^{n} I_S(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i) - \sum_{i=1}^{n} E(I_S(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)) \right| \leqslant \\ \leqslant \sqrt{\frac{n}{2} \ln \frac{2}{\delta}}. \end{aligned} \tag{3.36}$$

By definition of the deterministic forecast, the sums:

$$\sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - p_i) \text{ and } \sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - v^1)$$

differ at most by Δ for all i, where $v = (v^1, v^2)$. Summing (3.35) over $i = 1, \ldots, n$ and using the inequality (3.34), we obtain:

$$\begin{aligned} &\left| \sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)) \right| = \\ &= \left| \sum_{i=1}^{n} \sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - v^1) \right| \leqslant \\ &\leqslant \Delta n + 2\sqrt{(c_{\mathcal{F}}^2 + 1) n / \Delta^2} \end{aligned} \tag{3.37}$$

for all n.

By (3.36) and (3.37), for any S and n, with Pr-probability $1-\delta$:

$$\left|\sum_{i=1}^{n} I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)\right| \leqslant$$

$$\leqslant \Delta n + 2\sqrt{(c_{\mathcal{F}}^2+1)n/\Delta^2} + \sqrt{\frac{n}{2}\ln\frac{2}{\delta}}. \tag{3.38}$$

By Cauchy–Schwarz inequality:

$$\left|\sum_{n=1}^{N} D(x_n)(S_n - p_n)\right| =$$

$$= \left|\sum_{n=1}^{N} (S_n - p_n)(D \cdot \Phi(x_n))\right| =$$

$$= \left|\left(\sum_{n=1}^{N} (S_n - p_n)\Phi(x_n) \cdot D\right)\right| \leqslant$$

$$\leqslant \left\|\sum_{n=1}^{N} (S_n - p_n)\Phi(x_n)\right\|_{\mathcal{F}} \cdot \|D\|_{\mathcal{F}} \leqslant$$

$$\leqslant \|D\|_{\mathcal{F}}\sqrt{(c_{\mathcal{F}}^2+1)N}.$$

The proposition is proved. $\triangle$

Now we finish the proof of Theorem 3.5.

The expression $\Delta n + \sqrt{(c_{\mathcal{F}}^2+1)n/\Delta^2}$ from (3.37) and (3.38) takes its minimal value for $\Delta = \sqrt{2}(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} n^{-\frac{1}{4}}$. In this case, the right-hand side of the inequality (3.37) is equal to:

$$\Delta n + 2\sqrt{n(c_{\mathcal{F}}^2+1)/\Delta^2} = 2\Delta n = 2\sqrt{2}(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} n^{\frac{3}{4}}. \tag{3.39}$$

In what follows we use the upper bound $2\Delta n$ in (3.37).

To prove the bound (3.25) choose a monotonic sequence of rational numbers: $\Delta_1 > \Delta_2 > \ldots$ such that $\Delta_s \to 0$ as $s \to \infty$. We also define an increasing sequence of positive integer numbers: $n_1 < n_2 < \ldots$

For any s, we use for randomization on steps $n_s \leqslant n < n_{s+1}$ the partition of $[0,1]$ on subintervals of length Δ_s.

We start our sequences from $n_1 = 1$ and $\Delta_1 = 1$. Also, define the numbers $n_2, n_3, \ldots$ such that the inequality:

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i))\right| \leqslant 4(s+1)\Delta_s n \tag{3.40}$$

holds for all $n_s \leqslant n \leqslant n_{s+1}$ and for all $s \geqslant 1$.

We define this sequence by mathematical induction on s. Suppose that n_s ($s \geqslant 1$) is defined such that the inequality:

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i))\right| \leqslant 4s\Delta_{s-1} n \tag{3.41}$$

holds for all $n_{s-1} \leqslant n \leqslant n_s$, and the inequality:

$$\left|\sum_{i=1}^{n_s} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i))\right| \leqslant 4s\Delta_s n_s \tag{3.42}$$

also holds.

Let us define n_{s+1}. Consider all forecasts $\tilde{p}_i$ defined by the algorithm given above for the discretization $\Delta = \Delta_{s+1}$. We do not use first n_s of these forecasts (more correctly we will use them only in bounds (3.43) and (3.44); denote these forecasts $\hat{\mathbf{p}}_\mathbf{1}, \ldots, \hat{\mathbf{p}}_{\mathbf{n_s}}$). We add the forecasts $\tilde{p}_i$ for $i > n_s$ to the forecasts defined before this step of induction (for n_s). Let n_{s+1} be such that the inequality:

$$\begin{aligned}\left|\sum_{i=1}^{n_{s+1}} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i))\right| &\leqslant \left|\sum_{i=1}^{n_s} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i))\right| + \\ + \left|\sum_{i=n_s+1}^{n_{s+1}} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i)) + \sum_{i=1}^{n_s} E(I(\hat{\mathbf{p}}_\mathbf{i}, \tilde{x}_i)(y_i - \hat{\mathbf{p}}_\mathbf{i}))\right| + \\ + \left|\sum_{i=1}^{n_s} E(I(\hat{\mathbf{p}}_\mathbf{i}, \tilde{x}_i)(y_i - \hat{\mathbf{p}}_\mathbf{i}))\right| &\leqslant 4(s+1)\Delta_{s+1} n_{s+1}\end{aligned} \tag{3.43}$$

holds. Here the first sum of the right-hand side of the inequality (3.43) is bounded by $4s\Delta_s n_s$ – by the induction hypothesis (3.42). The

second and third sums are bounded by $2\Delta_{s+1}n_{s+1}$ and by $2\Delta_{s+1}n_s$, respectively, where $\Delta = \Delta_{s+1}$ is defined such that (3.39) holds. This follows from (3.37) and by choice of n_s.

The induction hypothesis (3.42) is valid for

$$n_{s+1} \geqslant \frac{2s\Delta_s + \Delta_{s+1}}{\Delta_{s+1}(2s+1)} n_s.$$

Similarly,

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i))\right| \leqslant \left|\sum_{i=1}^{n_s} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i))\right| +$$
$$+ \left|\sum_{i=n_s+1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i)) + \sum_{i=1}^{n_s} E(I(\hat{\mathbf{p}}_\mathbf{i}, \tilde{x}_i)(y_i - \hat{\mathbf{p}}_\mathbf{i}))\right| +$$
$$+ \left|\sum_{i=1}^{n_s} E(I(\hat{p}_i, \tilde{x}_i)(y_i - \hat{\mathbf{p}}_\mathbf{i}))\right| \leqslant 4(s+1)\Delta_s n \qquad (3.44)$$

for $n_s < n \leqslant n_{s+1}$. Here the first sum of the right-hand inequality (3.43) is also bounded: $4s\Delta_s n_s \leqslant 4s\Delta_s n$ – by the induction hypothesis (3.42). The second and the third sums are bounded by $2\Delta_{s+1}n \leqslant 2\Delta_s n$ and by $2\Delta_{s+1}n_s \leqslant 2\Delta_s n$, respectively. This follows from (3.37) and from choice of Δ_s. The induction hypothesis (3.41) is valid.

By (3.40) for any s

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(y_i - \tilde{p}_i))\right| \leqslant 4(s+1)\Delta_s n \qquad (3.45)$$

for all $n \geqslant n_s$ if Δ_s satisfies the condition $\Delta_{s+1} \leqslant \Delta_s(1 - \frac{1}{s+2})$ for all s.

We show now that sequences n_s and Δ_s satisfying all the conditions above exist.

Let $\epsilon > 0$ and $M = \lceil 2/\epsilon \rceil$, where $\lceil r \rceil$ is the least integer number greater than or equal to r. Define $n_s = (s+M)^M$ and $\Delta_s = \sqrt{2}(c_{\mathcal{F}}^2 + 1)^{\frac{1}{4}} n_s^{-\frac{1}{4}}$. Easy to verify that all requirements for n_s and Δ_s given above are satisfied for all $s \geqslant s_0$, where s_0 is sufficiently large. We

redefine $n_i = n_{s_0}$ for all $1 \leqslant i \leqslant s_0$. Note that these requirements hold for such i trivially.

We have in (3.41) for all $n_s \leqslant n < n_{s+1}$:

$$\begin{aligned} 4(s+1)\Delta_s n \leqslant 4(s+M)\Delta_s n_{s+1} = \\ = 4\sqrt{2}(c_{\mathcal{F}}^2+1)^{\frac{1}{4}}(s+M)(s+M+1)^M(s+M)^{-\frac{M}{4}} \leqslant \\ \leqslant 18(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} n_s^{\frac{3}{4}+2/M} \leqslant \\ \leqslant 18(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} n^{\frac{3}{4}+\epsilon}. \end{aligned}$$

Therefore, we obtain:

$$\left| \sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)) \right| \leqslant 18(c_{\mathcal{F}}^2+1)^{1/4} n^{3/4+\epsilon} \tag{3.46}$$

for all n.

Azuma–Hoeffding inequality says that for any $\gamma > 0$:

$$Pr\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} V_i \right| > \gamma \right\} \leqslant 2e^{-2n\gamma^2} \tag{3.47}$$

for all n, where V_i are martingale–differences (see Lemma 8.2 of Section 8.6).

We get $V_i = I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i) - E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))$ and $\gamma = \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$, where $\delta > 0$.

Combining (3.46) with (3.47), we obtain that for any $\delta > 0$, S and n, with probability $1 - \delta$:

$$\left| \sum_{i=1}^{n} I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i) \right| \leqslant 18(c_{\mathcal{F}}^2+1)^{1/4} n^{3/4+\epsilon} + \sqrt{\frac{n}{2} \ln \frac{2}{\delta}}.$$

Theorem 3.5 is proved. $\triangle$

3.5.2. Proof of Theorem 3.4

At any step i we compute the deterministic forecast p_i defined in Section 3.5.1 and its randomization to $\tilde{p}_i$ using parameters $\Delta = \Delta_s =$

$\sqrt{2}(c_{\mathcal{F}}+1)^{\frac{1}{4}}(s+M)^{-\frac{h}{4}}$ and $n_s = (s+M)^M$, where $n_s \leqslant i < n_{s+1}$. Let also, $\tilde{S}_{i-1}$ be a randomization of the past price S_{i-1}. In Theorem 3.5, $z_i = S_{i-1}$ and $\tilde{z}_i = \tilde{S}_{i-1}$.

The following upper bound directly follows from the method of discretization:

$$\begin{aligned}\left|\sum_{i=1}^{n} I(\tilde{p}_i > \tilde{S}_{i-1})(\tilde{S}_{i-1} - S_{i-1})\right| \leqslant \\ \leqslant \sum_{t=0}^{s}(n_{t+1} - n_t)\Delta_t \leqslant \\ \leqslant 4(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} n_s^{\frac{3}{4}+\epsilon} \leqslant \\ \leqslant 4(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} n^{\frac{3}{4}+\epsilon}, \qquad (3.48)\end{aligned}$$

where $n_s \leqslant n < n_{s+1}$.

Let $D(x)$ be an arbitrary trading strategy from RKHS $\mathcal{F}$. Clearly, the bound (3.48) holds if we replace $I(\tilde{p}_i > \tilde{S}_{i-1})$ on $\|D\|_\infty^{-1} D(\mathbf{x}_i)$.

For simplicity, we give the proof for the case of going long, where $D(x) \geqslant 0$ for all x and

$$\tilde{M}_i = \begin{cases} 1 \text{ if } \tilde{p}_i > \tilde{S}_{i-1}, \\ 0 \text{ otherwise.} \end{cases}$$

We use abbreviations:

$$\nu_1(n) = 4(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} n^{\frac{3}{4}+\epsilon}, \qquad (3.49)$$

$$\nu_2(n) = 18 n^{\frac{3}{4}+\epsilon}(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} + \sqrt{\frac{n}{2}\ln\frac{2}{\delta}}. \qquad (3.50)$$

$$\nu_3(n) = \sqrt{(c_{\mathcal{F}}^2+1)n} \qquad (3.51)$$

All sums below are for $i = 1, \dots n$. We use below the Azuma–Hoeffding inequality (3.47).

For any $\delta > 0$, with probability $1-\delta$ the following chain of equal-

ities and inequalities is valid:

$$\sum_{i=1}^{n} \tilde{M}_i(S_i - S_{i-1}) = \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (S_i - S_{i-1}) =$$

$$= \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (S_i - \tilde{p}_i) + \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (\tilde{p}_i - \tilde{S}_{i-1}) + \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (\tilde{S}_{i-1} - S_{i-1}) \geqslant \quad (3.52)$$

$$\geqslant \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (\tilde{p}_i - \tilde{S}_{i-1}) - \nu_1(n) - \nu_2(n) \geqslant \quad (3.53)$$

$$\geqslant \|D\|_\infty^{-1} \sum_{i=1}^{n} D(x_i)(\tilde{p}_i - \tilde{S}_{i-1}) -$$

$$-\nu_1(n) - \nu_2(n) =$$

$$= \|D\|_\infty^{-1} \sum_{i=1}^{n} D(x_i)(p_i - S_{i-1}) +$$

$$+\|D\|_\infty^{-1} \sum_{i=1}^{n} D(x_i)(\tilde{p}_i - p_i) -$$

$$-\|D\|_\infty^{-1} \sum_{i=1}^{n} D(x_i)(\tilde{S}_{i-1} - S_{i-1}) -$$

$$-\nu_1(n) - \nu_2(n) \geqslant \quad (3.54)$$

$$\geqslant \|D\|_\infty^{-1} \sum_{i=1}^{n} D(x_i)(p_i - S_{i-1}) -$$

$$-3\nu_1(n) - \nu_2(n) \geqslant \quad (3.55)$$

$$= \|D\|_\infty^{-1} \sum_{i=1}^{n} D(x_i)(S_i - S_{i-1}) -$$

$$-\|D\|_\infty^{-1} \sum_{i=1}^{n} D(x_i)(S_i - p_i) -$$

$$-3\nu_1(n) - \nu_2(n) = \quad (3.56)$$

$$= \|D\|_\infty^{-1} \sum_{i=1}^{n} D(x_i)(S_i - S_{i-1}) -$$

$$-3\nu_1(n) - \nu_2(n) - \|D\|_\infty^{-1} \|D\|_{\mathcal{F}} \nu_3(n).$$

In transition from (3.52) to (3.53) the inequality (3.25) of Theorem 3.5 and the bound (3.48) were used, and so, terms (3.49) and (3.50) were subtracted. In transition from (3.54) to (3.55) the bound (3.48) was applied twice to intermediate terms, and so, the term (3.48) was subtracted twice. In transition from (3.55) to (3.56) the inequality (3.26) of Theorem 3.5 has used, and so, the term (3.51) was subtracted.

Therefore, a constant $c > 0$ exists such that for any n and D, with probability $1 - \delta$,

$$\mathcal{K}_n^M \geqslant \|D\|^{-1} \mathcal{K}_n^D - c \left(n^{\frac{3}{4}+\epsilon} + (n \ln(1/\delta))^{\frac{1}{2}} \right). \tag{3.57}$$

The inequality (3.24) will follow from (3.57). For the proof we use the Borel–Cantelli lemma (see Section 8.6). This lemma states that if, for some sequence of events A_n the series $\sum\limits_{n=1}^{\infty} P(A_n)$ converges, then the probability that the event A_n holds for infinitely many n is 0.

In order to apply this lemma, we will return to the initial form of Hoeffding inequality. Denote $\gamma = \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$. Then $\delta = 2e^{-n\gamma^2}$. Rewrite (3.57) in the form:

$$\frac{1}{n} \mathcal{K}_n^M - \|D\|^{-1} \frac{1}{n} \mathcal{K}_n^D \geqslant -c \left(n^{-\frac{1}{4}+\epsilon} + \gamma \right) \tag{3.58}$$

According to (3.57), for any n and $\gamma > 0$, the inequality (3.58) violates with probability $2e^{-n\gamma^2}$. Since the series $\sum\limits_{n=1}^{\infty} e^{-n\gamma^2}$ converges, given γ the inequality (3.58) can be violated no more than for finite number of different n. By Borel–Cantelli lemma (see Section 8.6) the event:

$$\liminf_{n \to \infty} \frac{1}{n} \left(\mathcal{K}_n^M - \|f\|_{\infty}^{-1} \mathcal{K}_n^D \right) \geqslant 0$$

holds almost surely.

Theorem 3.4 is proved for any $D \in \mathcal{F}$.

Using a universal kernel and the corresponding canonical universal RKHS, we can extend our asymptotic results for all continuous stationary trading strategies D.

An RKHS $\mathcal{F}$ on X is universal if X is a compact metric space and every continuous function f on X can be arbitrarily well approximated in the metric $\|\cdot\|_\infty$ by a function from $\mathcal{F}$: for any $\epsilon > 0$ there exists $D \in \mathcal{F}$ such that

$$\sup_{x \in X} |f(x) - D(x)| \leqslant \epsilon$$

(see Steinwart [31], Definition 4).

We use $X = [0,1]$. The Sobolev space $\mathcal{F} = H^1([0,1])$ is the universal RKHS (see [31], [41]).

The existence of the universal RKHS on $[0,1]$ implies the full version of Theorem 3.4:

An algorithm for computing forecasts p_i and a sequential method of randomization can be constructed such that the randomized trading strategy $\tilde{M}_i$ performs at least as good as any nontrivial continuous trading strategy f:

$$\liminf_{n \to \infty} \frac{1}{n} \left(\mathcal{K}_n^M - \|f\|_\infty^{-1} \mathcal{K}_n^f \right) \geqslant 0 \tag{3.59}$$

holds almost surely with respect to a probability distribution generated by the corresponding sequential randomization.

This result directly follows from the inequality (3.57) and the possibility to approximate arbitrarily close any continuous function f on $[0,1]$ by a function D from the universal RKHS $\mathcal{F}$.

Any trading strategy $\tilde{M}_i$ satisfying (3.59) is called *universally consistent*.

The property of universal consistency (3.59) is strictly asymptotic and does not tell us anything about finite data sequences. We have obtained the convergence bound (3.57) for more narrow classes of functions like RKHS.

3.6. Problems

1. Prove that when using the method of Krichevsky and Trofimov the conditional probability of $\omega_{n+1} = 1$ given n binary observations

$\omega^n = \omega_1, \ \ldots, \ \omega_n$ equals

$$P(1|\omega^n) = \frac{n_1 + 1/2}{n+1}.$$

2. Prove that the following bound is valid:

$$\int_0^1 \frac{p^{n_1}(1-p)^{n_2}}{\pi\sqrt{p(1-p)}} dp \geqslant \frac{1}{2\sqrt{n}} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}.$$

3. For some sequences it is easy to construct calibrated predictions. A binary sequence $\omega_1, \omega_2, \ldots$ is called stationary if the limit

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \omega_i$$

exists. Prove that the sequence of forecasts $p_1, p_2, \ldots$ defined by $p_1 = 0$ and

$$p_i = \frac{1}{i-1} \sum_{j=1}^{i-1} \omega_j$$

for $i > 1$ calibrates on a stationary sequence $\omega_1, \omega_2, \ldots$.

4. Prove that no randomizing algorithm exists such that for any sequence of binary outcomes $\omega_1, \omega_2, \ldots$ a modified condition of calibration holds:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} |\omega_i - \tilde{p}_i| = 0$$

with probability one, where $\tilde{p}_i$ are the corresponding random forecasts.

3.7. Laboratory work

The algorithm described in Section 3.3, can be easily implemented as a computer program. In this case, to calculate a root of the equation (3.6) is better to use a smooth approximation to the kernel (3.5) the

Gaussian kernel $K(p,p') = e^{-\gamma(p-p')^2}$ for some $\gamma > 0$. [5] You can also use the kernel of the form $K(p,p') = \cos(\gamma(p-p'))$ A root of the equations (3.6) or (3.14) can be found by the method of sequential bisection of the unit interval.

Different time series can be downloaded from the site FINAM: http://old.finam.ru/analysis/export/default.asp

For example, you can download per-minute data of stock prices of some companies:

$$S_0, S_1, S_2, \ \ldots, \ S_n$$

and normalize them so that $S_i \in [0,1]$ for all i.

Laboratory work 1

Implement the algorithm Section 3.3. Write a program to compute well-calibrated forecasts $p_1, p_2, \ \ldots, \ p_n$ for a binary sequence $\omega_1, \omega_2, \ldots, \omega_n$, where $\omega_i \in \{0,1\}$. Compare these predictions with the predictions computed using the Laplace rule.

Laboratory work 2

Download the time series of prices of a stock. Normalize the stock price $S_0, S_1, \ \ldots, S_{n-1}$ so that $S_i \in [0,1]$.

Write a program to compute the well-calibrated forecasts $p_1, p_2, \ \ldots, \ p_n$ for the sequence of real numbers $S_0, S_1, \ \ldots, S_{n-1}$. Visualize the results.

Propose and implement the computer programs for trading with stock prices, using well-calibrated forecasts. Select sequences of prices using rules of the type $p_i > \omega_{i-1} + \delta$ for $\delta \geqslant 0$. Visualize the results.

[5] In this case signals are absent. To use signals, you can use the kernel (3.15).

Chapter 4

Prediction with Expert Advice

The problem of making the right rational decisions is central in science and practice. A decision is made on basis of some observations. As in the previous chapter, we consider the problem of predicting parameters of a process. Only now, we evaluate validity of our predictions guided by different principles. We also will not use any assumptions about the nature of mechanisms generating the observed data.

The correct forecast or the right decision leads to a smaller loss than a wrong decision. In the traditional statistical approach, we compare the loss suffered from our forecasts with some ideal model of decision-making, which is usually based on a statistical model describing the observed data. In the traditional approach, at first, we estimate parameters of our model, and, after that, we compute a forecast using this model.

At the competitive approach, instead of a single ideal model, a variety of possible models are considered. They are called expert strategies or simply experts. A set of such expert strategies can be finite or infinite and even uncountable. Using the outcomes received online, such an expert strategy produces predictions of the future outcomes. The learning algorithm observes the forecasts of these competing strategies and evaluate their performance in the past. The

learning algorithm makes its forecast using this evaluation.

The performance of our algorithm is compared with the performance of the experts. Usually, we compare the loss suffered by our algorithm over period of prediction with the loss suffered by the best expert algorithm.

A comparison can be made as in the worst case, as well as on the average when our forecasting algorithm uses randomization. Note that the internal probability distribution using by the randomized algorithm has no relation to a source generating outcomes. As we say, our algorithm uses a random number generator.

Let us discuss the types of processes generating data that will be predicted by our forecasting methods. The behavior of some processes is independent on the predictions issued by the forecaster. Such processes are often considered in classical mechanics and physics. For example, the weather is independent on the predictions of a weather forecaster.

The methods presented below work in same way in the case, when the parameters of the process depend on the predictions made by the forecaster. This is the so-called case of adaptive *adversatively* nature. For example, this assumption is natural for forecasting in financial games and social processes.

4.1. Weighted Majority Algorithm

In this section we consider the simplest algorithms for precise prediction of future outcomes. There are two possible outcomes 0 and 1. There are N experts (strategies) that at every step output the predictions $p_t^i \in \{0, 1\}$, $i = 1, \ldots, N$.

The algorithm's goal is to predict a future outcome of an infinite binary sequence $\omega_1, \ldots, \omega_{t-1}$ whose bits are revealed one at a time. A prediction p_t is correct if $p_t = \omega_t$, the algorithm make a mistake at step t otherwise.

Just before the tth bit is revealed, a set of N experts make predictions $p_t^1, \ldots, p_t^N$. The algorithm is allowed to observe all of these predictions, then it makes a guess $p_t \in \{0, 1\}$, and then the truth, ω_t, is revealed. We are given a promise that there is at least one expert

whose predictions are always accurate, ie, we are promised that an i exists such that $p_i^t = \omega_t$ for all t.

Consider the following algorithm, which is called the "Majority algorithm". At each time t, it consults the predictions of all experts who did not make a mistake during one of the t steps. In other words, it considers the set of experts

$$B_t = \{i : p_j^i = \omega_j \text{ for all } 1 \leqslant j \leqslant t-1\}$$

The majority algorithm outputs a forecast $p_t = 1$ if if at least half of experts predict 1, and outputs $p_t = 0$ otherwise:

$$p_t = \begin{cases} 1 \text{ if } |\{i : i \in B_t, p_t^i = 1\}| \geqslant |B_t|/2, \\ 0 \text{ otherwise.} \end{cases}$$

Theorem 4.1. *Assume that an expert i exists such that $p_t^i = \omega_t$ for all t. Then the "majority algorithm" makes no more than $\lceil \log_2 N \rceil$ mistakes, where N is the number of experts.*

Proof. If the "majority algorithm" makes a mistake at step t than at least half of the experts in B_t made a mistake at that time, so the number of previously never mistaken experts reduced by at least half: $|B_{t+1}| \leqslant \lceil |B_t|/2 \rceil$. By the assumption $|B_t| \geqslant 1$ for all t. Hence decrease in the value the number $|B_t|$ twice is at most $\lceil \log_2 N \rceil$. $\triangle$

Now consider the case where an expert, just guesses the future outcomes, does not exist. In this case, consider the "Weighted Majority Algorithm", which was discovered by Littlestone and Warmuth [22].

Let ϵ be a parameter such that $0 < \epsilon < 1$. Define $w_1^i = 1$ for all $1 \leqslant i \leqslant N$.

Algorithm $WMA(\epsilon)$
Define $w_1^i = 1$ for $i = 1, \ldots, N$.

FOR $t = 1, 2, \ldots, T$
Expert i announces a forecast $p_t^i \in \{0, 1\}$, $i = 1, \ldots, N$
Learner announces a forecast p_t of the algorithm $WMA(\epsilon)$:

IF $\sum\limits_{i:p_t^i=0} w_t^i > \sum\limits_{i:p_t^i=1} w_t^i$
THEN $p_t = 0$
ELSE $p_t = 1$
ENDIF
Nature announces an outcome $\omega_t \in \{0, 1\}$
Learner updates the expert weights:

Let $E_t = \{i : p_t^i \neq \omega_t\}$ be the set of experts i, who made a mistake at step t
Reduce the weights of such experts:

$$w_{t+1}^i = \begin{cases} (1-\epsilon)w_t^i \text{ if } i \in E_t, \\ w_t^i \text{ otherwise} \end{cases}$$

ENDFOR

Let $L_T^i = \sum\limits_{t=1}^{T} |p_t^i - \omega_t|$ be a number of all mistakes of *Expert* i and $L_T = \sum\limits_{t=1}^{T} |p_t - \omega_t|$ be a number of all mistakes of *Learner*, ie, of the algorithm $WMA(\epsilon)$ on the first T steps.

Theorem 4.2. *The number of mistakes of the $WMA(\epsilon)$ algorithm has a bound*

$$L_T \leqslant \left(\frac{2}{1-\epsilon}\right) \min_{1 \leqslant i \leqslant N} L_T^i + \left(\frac{2}{\epsilon}\right) \ln N$$

for all T.

Proof. Define $W_t = \sum\limits_{i=1}^{N} w_t^i$. Let $m = \min\limits_{1 \leqslant i \leqslant N} L_T^i$ be the number of mistakes of the best expert for T steps. Assume the best expert is i. Then the weight of the ith expert has updated at most m times. Then

$$W_t > w_t^i \geqslant (1-\epsilon)^m \tag{4.1}$$

for all t such that $1 \leqslant t \leqslant T$.

On the other hand, if the algorithm makes a mistake at step t then

$$\sum_{i \in E_t} w_t^i \geqslant W_t/2.$$

Hence,

$$\begin{aligned} W_{t+1} = \sum_{i \in E_t} (1-\epsilon) w_t^i + \sum_{i \notin E_t} w_t^i = \\ = \sum_{i=1}^{N} w_t^i - \epsilon \sum_{i \in E_t} w_t^i \leqslant \\ \leqslant W_t \left(1 - \frac{\epsilon}{2}\right). \end{aligned}$$

By definition $W_{t+1} \leqslant W_t$ for all t. Then for any $T > 0$,

$$\frac{W_T}{W_1} = \prod_{t=1}^{T-1} \frac{W_{t+1}}{W_t} \leqslant \left(1 - \frac{\epsilon}{2}\right)^M, \tag{4.2}$$

where $M = L_T$ is the total number of mistakes of the algorithm $WMA(\epsilon)$ for the first T steps.

By definition $W_1 = \sum\limits_{i=1}^{N} w_1^i = N$. From (4.1) and (4.2), we have

$$\frac{(1-\epsilon)^m}{N} < \frac{W_T}{W_1} \leqslant \left(1 - \frac{\epsilon}{2}\right)^M.$$

Now we take the natural logarithm of both sides of this inequality and make the following transformations:

$$\begin{aligned} m \ln(1-\epsilon) - \ln N < M \ln\left(1 - \frac{\epsilon}{2}\right) \\ m \ln(1-\epsilon) - \ln N < -\frac{\epsilon}{2} M \\ m \ln\left(\frac{1}{1-\epsilon}\right) + \ln N > \frac{\epsilon}{2} M \\ m \left(\frac{2}{\epsilon}\right) \ln\left(\frac{1}{1-\epsilon}\right) + \left(\frac{2}{\epsilon}\right) \ln N > M \\ \left(\frac{2}{1-\epsilon}\right) m + \left(\frac{2}{\epsilon}\right) \ln N > M, \end{aligned} \tag{4.3}$$

The second line of (4.3) was derived from the first one using inequality $\ln(1+x) \leqslant x$ that holds for all $x > -1$.

The last line of (4.3) was derived from the previous line using inequality

$$\frac{1}{y} \ln\left(\frac{1}{1-y}\right) \leqslant \frac{1}{1-y}.$$

This inequality can be derived from the inequality $\ln(1+x) \leqslant x$ substituting $x = y/(1-y)$. $\triangle$

Theorem 4.2 shows that the weighted majority algorithm errs no more than about two times greater than the best expert.

Historically, it was the first algorithm of this kind. It was proposed by Littlestone and Varmuth in 1989 and was called "Weighted Majority Algorithm" [22]. Later, in 1990, Vovk [35] proposed a more general algorithm "Aggregating Algorithm" and the concept of mixability that work for a more general type of games.

4.2. Algorithm for solving the dynamic allocation problem

In this section, we describe a simple algorithm for solving a dynamic allocation problem in case where only experts losses are known. This algorithm was proposed by Freund and Shapire [14].

Let us explain the idea of this algorithm by the following example. A gambler decides to allow a group of his fellow gamblers to make bets on his behalf. He decides he will wager a fixed sum of money in every race, but that he will apportion his money among his friends based on how well they are doing. Certainly, if he knew psychically ahead of time which of his friends would win the most, he would naturally have that friend handle all his wagers. Lacking such clairvoyance, however, he attempts to allocate each race's wager in such a way that his total winnings for the season will be reasonably close to what he would have won had he bet everything with the luckiest of his friends.

We formalize the online allocation model as follows. The forecasting process is presented in the form of a perfect-information game. Players are: strategies or *Experts* i, $i = 1, 2, \ldots, N$, and *Allocator*.

The goal of *Allocator* is to minimize its cumulative loss relative to the loss suffered by the best strategy.

At each step $t = 1, 2, \ \ldots, \ T$, *Allocator* decides on a distribution $\bar{p}_t = (p_t^1, \ \ldots, \ p_t^N)$ over the strategies, where $p_t^1 + \cdots + p_t^N = 1$ and $p_t^i \geqslant 0$ for $i = 1, 2, \ \ldots, \ N$.

Each strategy i then suffers some loss l_t^i at step t, where $i = 1, 2, \ \ldots, \ N$, which is determined by the (possibly adversarial) "environment."

The loss suffered at step t by *Allocator* is then the average loss of the strategies with respect to chosen allocation rule:

$$(\bar{p}_t \cdot \bar{l}_t) = \sum_{i=1}^{N} p_t^i l_t^i,$$

where $\bar{l}_t = (l_t^1, \ \ldots, \ l_t^N)$ is the vector of losses suffered by all strategies on step t. We call this loss function the mixture loss.

We assume that the loss suffered by any strategy is bounded so that, without loss of generality, $l_t^i \in [0, 1]$ for all i and t.

Besides this condition, we make no assumptions about the form of the loss vectors l_t^i, or about the manner in which they are generated; indeed, the adversary's choice for l_t^i may even depend on the allocator's chosen mixture $\bar{p}_t$.

In the case of limited losses at each step there is no fundamental difference between the algorithms that achieve the minimum loss, and algorithms, that achieve the maximum payoff. We can move from losses l_t at each step t to gain $1 - l_t$, and vise wersa.

The cumulative loss of *Expert* i for steps $t = 1, 2, \ \ldots, \ T$ equals

$$L_T^i = \sum_{t=1}^{T} l_t^i.$$

Accordingly, the cumulative loss of *Allocator* for steps $t = 1, 2, \ \ldots, \ T$ equals

$$L_T = \sum_{t=1}^{T} (\bar{p}_t \cdot \bar{l}_t).$$

The goal of *Allocator* is to develop a strategy $\bar{p}_t$, $t = 1, 2, \ldots, T$, which minimizes its cumulative loss relative to the loss suffered by the best strategy. That is, *Allocator* attempts to minimize its net loss

$$R_T = L_T - \min_i L_T^i.$$

We show that Littlestone and Warmuth's "weighted majority" algorithm can be generalized to handle this problem, and we prove a number of bounds on the net loss. The problem can be solved by the algorithm $Hedge(\beta)$ proposed by Freund and Shapire [14].

Parameters of this algorithm are a real number $\beta \in (0, 1)$ and a vector of initial weights $\bar{w}_1 = (w_1^1, \ldots, w_1^N)$.

Assume that the initial weights of all experts satisfies the equality $\sum_{i=1}^{N} w_1^i = 1$.

The online allocation algorithm $Hedge(\beta)$
FOR $t = 1, 2, \ldots, T$
Allocator computes distribution of the expert strategies:

$$\bar{p}_t = \frac{\bar{w}_t}{\sum_{i=1}^{N} w_t^i}. \tag{4.4}$$

Expert i suffers its loss l_t^i, where $i = 1, 2, \ldots, N$. Denote $\bar{l}_t = (l_t^1, \ldots, l_t^N)$ the vector of the expert losses at step t.
Allocator suffers its loss: $l_t = (\bar{p}_t \cdot \bar{l}_t)$.
Allocator updates weights of the experts:

$$w_{t+1}^i = w_t^i \beta^{l_t^i} \tag{4.5}$$

for $i = 1, \ldots, N$.
ENDFOR

Lemma 4.1. *For any sequence of vectors $\bar{l}_1, \ldots, \bar{l}_T$ of the experts losses, the following inequality holds:*

$$\ln\left(\sum_{i=1}^{N} w_{T+1}^i\right) \leqslant -(1-\beta)L_T, \tag{4.6}$$

where L_T is a loss of Allocator for the first T steps.

Proof. By a convexity argument $\beta^r \leqslant 1-(1-\beta)r$ for all $r \in [0,1]$ and $0 < \beta < 1$. Combining this inequality with (4.4) and (4.5), we obtain:

$$\sum_{i=1}^{N} w_{t+1}^i = \sum_{i=1}^{N} w_t^i \beta^{l_t^i} \leqslant$$

$$\leqslant \sum_{i=1}^{N} w_t^i (1-(1-\beta)l_t^i) =$$

$$= \left(\sum_{i=1}^{N} w_t^i\right)(1-(1-\beta)(\bar{p}_t \cdot \bar{l}_t)). \tag{4.7}$$

Applying repeatedly (4.7) for $t = 1, ..., T$, we obtain

$$\sum_{i=1}^{N} w_{T+1}^i \leqslant$$

$$\leqslant \prod_{t=1}^{T}(1-(1-\beta)(\bar{p}_t \cdot \bar{l}_t)) \leqslant$$

$$\leqslant \exp\left(-(1-\beta)\sum_{t=1}^{T}(\bar{p}_t \cdot \bar{l}_t)\right).$$

We have used here the inequality $1 + x \leqslant \exp(x)$ for all x and the equality $\sum\limits_{i=1}^{N} w_1^i = 1$. Lemma is proved. $\triangle$

By (4.6) we have:

$$L_T \leqslant \frac{-\ln\left(\sum\limits_{i=1}^{N} w_{T+1}^i\right)}{1-\beta}. \tag{4.8}$$

By definition of weights (4.5)

$$w_{T+1}^i = w_1^i \prod_{t=1}^{T} \beta^{l_t^i} = w_1^i \beta^{L_T^i}. \tag{4.9}$$

Hence we obtain the following theorem.

Theorem 4.3. *For any sequence of vectors* $\bar{l}_1, \ \ldots, \ \bar{l}_T$ *of the experts losses and for any* i*:*

$$L_T \leqslant \frac{-\ln(w_1^i) - L_T^i \ln \beta}{1-\beta}. \tag{4.10}$$

In the case of a finite number of experts it is natural to set the initial weights of expert strategies equal to $w_1^i = \frac{1}{N}$ for all i. Then (4.10) can be written in the form:

$$L_T \leqslant \frac{\ln(1/\beta)}{1-\beta} \min_i L_T^i + \frac{\ln N}{1-\beta}. \tag{4.11}$$

Inequality (4.11) can be interpreted as the fact that the cumulative loss of allocation algorithm $Hedge(\beta)$ does not exceed the loss of the best expert, multiplied by the constant $\frac{\ln(1/\beta)}{1-\beta}$ plus "the regret" $\frac{\ln N}{1-\beta}$.

Since the regret depends only logarithmical on N, this bound is reasonable even for a very large number of strategies.

Vovk [36] analyzed prediction algorithms that have performance bounds of this form, and proved the tight upper and lower bounds for the achievable values of c and a. Using Vovk's results, one can show that the constants a and c achieved by $Hedge(\beta)$ are optimal.

Theorem 4.4. *Let* B *be an arbitrary allocation algorithm working with any finite number of experts.*

Assume that positive real numbers a *and* c *exist such that for each* N *strategies and for each sequence of expert losses* $\bar{l}^1, \ \ldots, \ \bar{l}^T$*, where* $\bar{l}^t = (l_1^t, \ \ldots, \ l_N^t)$ *for* $t = 1, \ \ldots, \ T$*, the following inequality holds:*

$$L_T(B) \leqslant c \min_i L_T^i + a \ln N.$$

Then for all $\beta \in (0,1)$ *will be one of the inequalities:*

$$c \geqslant \frac{\ln(1/\beta)}{1-\beta} \ \text{ or } a \geqslant \frac{1}{1-\beta}.$$

In practice, we will often want to choose β so as to maximally exploit any prior knowledge we may have about the specific problem at hand.

By choosing the parameter β, it is possible to achieve the redistribution of the constants so that a multiplicative factor in (4.11) became equal to one due to increase the additive term.

The following lemma will be helpful for choosing β using the bounds derived above.

Lemma 4.2. *Assume that* $0 \leqslant L \leqslant \tilde{L}$ *and* $0 \leqslant R \leqslant \tilde{R}$. *Let also,* $\beta = g(\tilde{L}/\tilde{R})$, *where*

$$\beta = \frac{1}{1 + \sqrt{\frac{2\tilde{R}}{\tilde{L}}}}.$$

Then

$$-\frac{\ln \beta}{1-\beta} L + \frac{1}{1-\beta} R \leqslant L + \sqrt{2\tilde{L}\tilde{R}} + R. \tag{4.12}$$

Proof. We use the following well known inequality: $-\ln \beta \leqslant \frac{1-\beta^2}{2\beta}$ for $\beta \in (0, 1]$. The following chain of transformations leads to the desired result:

$$\begin{aligned}
L\frac{-\ln \beta}{1-\beta} + \frac{1}{1-\beta} R &\leqslant L\frac{1+\beta}{2\beta} + \frac{1}{1-\beta} R = \\
&= \frac{1}{2} L \left(1 + \frac{1}{\beta}\right) + \frac{1}{1-\beta} R = \\
&= L + \frac{1}{2} L \sqrt{\frac{2\tilde{R}}{\tilde{L}}} + \frac{1}{1 - \frac{1}{1+\sqrt{\frac{2\tilde{R}}{\tilde{L}}}}} R \leqslant \\
&\leqslant L + \sqrt{\frac{1}{2}\tilde{L}\tilde{R}} + R + R\sqrt{\frac{\tilde{L}}{2\tilde{R}}} \leqslant \\
&\leqslant L + \sqrt{2\tilde{L}\tilde{R}} + R.
\end{aligned}$$

Since we assumed that $0 \leqslant l_t^i \leqslant 1$ for all i and t, the cumulative losses of all experts are bounded: $L_T^i \leqslant T$ for all i and T. Therefore, we can take $\tilde{L} = T$ in (4.12). Take also $\tilde{R} = \ln N$. Then by Lemma 4.2:

$$L_T \leqslant \min_i L_T^i + \sqrt{2T \ln N} + \ln N, \tag{4.13}$$

where L_T is the cumulative loss of the algorithm $Hedge(\beta)$ over first T steps.

The drawback of this estimate is that the parameter β depends on the horizon T. See also the comment at the end of Section 4.4.

The bound given in (4.13) can be improved in special cases in which the loss is a function of a prediction and an outcome and this function is of a special form However, it possible to prove that for the general case, one cannot improve the square-root term $\sqrt{2T \ln N}$ by more than a constant factor.

4.3. Follow the perturbed leader

In this section we consider a different general approach— "Follow the Perturbed Leader – FPL" algorithm, now called Hannan's algorithm, see Hannan [15], Kalai and Vempala [19] and Lugosi and Cesa-Bianchi [23]. Hutter and Poland [16] presented a further developments of the FPL algorithm for a countable class of experts, arbitrary weights and adaptive learning rate.

Under this approach we only choose the decision that has fared the best in the past—the leader. In order to cope with adversary some randomization is implemented by adding a perturbation to the total loss prior to selecting the leader. The goal of the learner's algorithm is to perform almost as well as the best expert in hindsight in the long run. The resulting FPL algorithm has almost the same performance guarantees as WM-type algorithms for fixed learning rate and bounded one-step losses.

Prediction with Expert Advice considered in this section proceeds as follows. We are asked to perform sequential actions at times $t = 1, 2, \ldots, T$. At each time step t, experts $i = 1, \ldots N$ receive results of their actions in form of their losses s_t^i—arbitrary real numbers.

At the beginning of the step t *Learner*, observing cumulating losses $s_{1:t-1}^i = s_1^i + \cdots + s_{t-1}^i$ of all experts $i = 1, \ldots N$, makes a decision to follow one of these experts, say Expert i. At the end of step t *Learner* receives the same loss s_t^i as Expert i at step t and suffers *Learner's* cumulative loss $s_{1:t} = s_{1:t-1} + s_t^i$.

We suppose that one-step losses of all experts are bounded, for

example, $0 \leqslant s_t^i \leqslant 1$ for all i and t.

Well known simple example of a game with two experts shows that *Learner* can perform much worse than each expert: let the current losses of two experts on steps $t = 0, 1, \ldots, 6$ be $s^1_{0,1,2,3,4,5,6} = (\frac{1}{2}, 0, 1, 0, 1, 0, 1)$ and $s^2_{0.1,2,3,4,5,6} = (0, 1, 0, 1, 0, 1, 0)$. Evidently, "Follow the Leader" algorithm always chooses the wrong prediction.

When the experts one-step losses are bounded, this problem has been solved using randomization of the experts cumulative losses and only then we choose the best expert.

The FPL algorithm outputs prediction of an expert i which minimizes

$$s^i_{1:t-1} - \frac{1}{\epsilon}\xi^i,$$

where ϵ is *a learning rate*, and ξ^i, $i = 1, \ldots N$, $t = 1, 2, \ldots$, is a sequence of i.i.d. nonnegative random variables distributed according to the exponential distribution with the density $p(x) = \exp\{-x\}$, $x \geqslant 0$.

We use the properties of this distribution: $P\{\xi > a\} = e^{-a}$ and $P\{\xi > a + b\} = e^{-b}P\{\xi > a\}$ for all nonnegative a and b. We refer a reader for problems in Section 4.8 for a proof.

At each step t of the game, all N experts receive one-step losses $s^i_t \in [0, 1]$, $i = 1, \ldots N$, and the cumulative loss of the ith expert after step t is equal to

$$s^i_{1:t} = s^i_{1:t-1} + s^i_t.$$

Assume that $\epsilon_t = a/\sqrt{t}$ for all t, where a constant a will be specified below. We suppose without loss of generality that $s^i_0 = v_0 = 0$ for all i and $\epsilon_0 = \infty$.

The FPL algorithm is defined on Figure 4.1.

Let $s_{1:T} = \sum_{t=1}^{T} s_t^{I_t}$ be the cumulative loss of the FPL algorithm for first T steps.

The following theorem presents an upper bound for the regret of the FPL algorithm.

We suppose that the experts are oblivious, that is, they do not use in their work random actions of the learning algorithm.

FPL algorithm.

FOR $t = 1, \dots T$
Learner chooses an expert with the minimal perturbed cumulated loss on past steps $< t$:

$$I_t = \text{argmin}_{i=1,2,\dots N} \left\{ s^i_{1:t-1} - \frac{1}{\epsilon_t} \xi^i \right\}.$$

Expert i suffers a loss s^i_t for $i = 1, \dots, N$.
Learner suffers the loss $s_t = s^{I_t}_t$.
ENDFOR

Figure 4.1: FPL algorithm

Theorem 4.5. *The expected cumulated loss of the* FPL *algorithm with the variable learning rate* $\epsilon_t = \sqrt{\frac{2 \ln N}{t}}$ *has the bound:*

$$E(s_{1:T}) \leqslant \min_i s^i_{1:T} + 2\sqrt{2T \ln N} \tag{4.14}$$

The algorithm FPL *is asymptotically consistent:*

$$\limsup_{T \to \infty} \frac{1}{T} (s_{1:T} - \min_{i=1,\dots N} s^i_{1:T}) \leqslant 0 \tag{4.15}$$

with probability one.

Proof. The analysis of optimality of the FPL algorithm is based on an intermediate predictor IFPL (Infeasible FPL) (see Figure 4.2).

The IFPL algorithm predicts under the knowledge of $s^i_{1:t}$, $i = 1, \dots N$, which may not be available at beginning of step t.

The expected one-step and cumulated losses of the FPL and IFPL algorithms at steps t and T are denoted:

$$l_t = E(s^{I_t}_t) \text{ and } r_t = E(s^{J_t}_t),$$
$$l_{1:T} = \sum_{t=1}^{T} l_t \text{ and } r_{1:T} = \sum_{t=1}^{T} r_t,$$

IFPL algorithm.

FOR $t = 1, \dots T$
Learner chooses an expert with the minimal perturbed cumulated loss on steps $\leqslant t$:

$$J_t = \mathrm{argmin}_{i=1,2,\dots N} \left\{ s^i_{1:t} - \frac{1}{\epsilon_t}\xi^i \right\}.$$

Expert i suffers a loss s^i_t for $i = 1, \dots, N$.
Learner suffers the loss $s^{J_t}_t$.
ENDFOR

Figure 4.2: IFPL algorithm

respectively, where $s^{I_t}_t$ is the one-step loss of the FPL algorithm at step t and $s^{J_t}_t$ is the one-step loss of the IFPL algorithm, and E denotes the mathematical expectation. Recall that $I_t = \mathrm{argmin}_i\{s^i_{1:t-1} - \frac{1}{\epsilon_t}\xi^i\}$ and $J_t = \mathrm{argmin}_i\{s^i_{1:t} - \frac{1}{\epsilon'_t}\xi^i\}$.

Lemma 4.3. *The expected cumulated losses of the* FPL *and* IFPL *algorithms satisfy the inequality:*

$$l_{1:T} \leqslant r_{1:T} + \sum_{t=1}^{T} \epsilon_t \tag{4.16}$$

for all T.

Proof. Let $c_1, \dots c_N$ be arbitrary nonnegative real numbers. For any $1 \leqslant j \leqslant N$, define the numbers m_j and m'_j:

$$\begin{aligned} m_j = \min_{i \neq j}\{s^i_{1:t-1} - \frac{1}{\epsilon_t}c_i\} \leqslant \\ \leqslant \min_{i \neq j}\{s^i_{1:t-1} + s^i_t - \frac{1}{\epsilon_t}c_i\} = \\ = \min_{i \neq j}\{s^i_{1:t} - \frac{1}{\epsilon_t}c_i\} = m'_j. \end{aligned}$$

Comparing conditional probabilities:

$$P\{I_t = j | \xi^i = c_i, i \neq j\} \text{ and } P\{J_t = j | \xi^i = c_i, i \neq j\}$$

is the core of the proof of the lemma. It holds:

$$\begin{aligned}
P\{I_t = j|\xi^i = c_i, i \neq j\} = \\
= P\{s^j_{1:t-1} - \frac{1}{\epsilon_t}\xi^j \leqslant m_j|\xi^i = c_i, i \neq j\} = \\
= P\{\xi^j \geqslant \epsilon_t(s^j_{1:t-1} - m_j)|\xi^i = c_i, i \neq j\} \leqslant \\
\leqslant e^{\epsilon_t} P\{\xi^j \geqslant \epsilon_t(s^j_{1:t-1} - m_j + 1)|\xi^i = c_i, i \neq j\} \leqslant \\
\leqslant e^{\epsilon_t} P\{\xi^j \geqslant \epsilon_t(s^j_{1:t-1} + s^i_t - m_j)|\xi^i = c_i, i \neq j\} \leqslant \\
\leqslant e^{\epsilon_t} P\{\xi^j \geqslant \epsilon_t(s^j_{1:t} - m'_j)|\xi^i = c_i, i \neq j\} = \\
= e^{\epsilon_t} P\{s^j_{1:t} - \frac{1}{\epsilon_t}\xi^j \leqslant m'_j|\xi^i = c_i, i \neq j\} = \\
= e^{\epsilon_t} P\{J_t = j|\xi^i = c_i, i \neq j\}. \qquad (4.17)
\end{aligned}$$

We have used in transition from the 3th row to the 4th row the inequality $P\{\xi \geqslant a + b\} \leqslant e^{-b}P\{\xi \geqslant a\}$ for any random variable ξ distributed according to the exponential law, where a and b are arbitrary nonnegative real numbers.

Since this bound holds under any conditions c_i, it also holds unconditionally:

$$P\{I_t = j\} \leqslant e^{\epsilon_t} P\{J_t = j\}. \qquad (4.18)$$

for all $t = 1, 2, \ldots$ and $j = 1, \ldots N$.

Summing (4.18) over $t = 1\ , \ldots,\ T$, we obtain the inequality for expectation of one-step losses:

$$l_t = E(s^{I_t}_t) = \sum_{j=1}^{T} s^j_t P\{I_t = j\} \leqslant e^{\epsilon_t} \sum_{j=1}^{T} s^j_t P\{J_t = j\} = e^{\epsilon_t} r_t.$$

Finally, $l_t - r_t \leqslant \epsilon_t l_t$ follows from $r_t \geqslant e^{-r} l_t \geqslant (1 - r) l_t$ for $r \leqslant 1$. Summing over $t = 1\ , \ldots,\ T$ and taking into account that $0 \leqslant l_t \leqslant 1$ for all t, we obtain:

$$l_{1:T} \leqslant r_{1:T} + \sum_{t=1}^{T} \epsilon_t \leqslant r_{1:T} + 2a\sqrt{T}.$$

Lemma is proved. $\triangle$

The following lemma gives a bound for the expected cumulative loss of the IFPL algorithm.

Lemma 4.4. *The expected cumulative loss of the* IFPL *algorithm has the bound:*

$$r_{1:T} \leqslant \min_i s^i_{1:T} + \frac{\ln N}{\epsilon_T} \tag{4.19}$$

for all T.

Proof. Let in this proof, $\mathbf{s}_t = (s^1_t, \dots s^N_t)$ be a vector of one-step losses and $\mathbf{s}_{1:t} = (s^1_{1:t}, \dots s^N_{1:t})$ be a vector of cumulative losses of the experts algorithms. Also, let $\xi = (\xi^1, \dots \xi^N)$ be a vector whose coordinates are exponentially distributed random variables.

Consider the auxiliary vectors of modified losses:

$$\tilde{\mathbf{s}}_t = \mathbf{s}_t - \xi\left(\frac{1}{\epsilon_t} - \frac{1}{\epsilon_{t-1}}\right) \tag{4.20}$$

$$\tilde{\mathbf{s}}_{1:t} = \mathbf{s}_{1:t} - \frac{1}{\epsilon_t}\xi \tag{4.21}$$

for $t = 1, 2, \dots$ for the moment.

For any vector $\mathbf{s} = (s^1, \ \dots, \ s^N)$ and a unit vector $\mathbf{d} = (0, \ \dots, \ 1, \ \dots, \ 0)$, denote

$$M(\mathbf{s}) = \mathrm{argmin}_{\mathbf{d}\in D}\{\mathbf{d}\cdot\mathbf{s}\},$$

where $D = \{(0, \ \dots, \ 1), \ \dots, \ (1, \ \dots, \ 0)\}$ is the set of N unit vectors of dimension N and "$\cdot$" is the dot product of two vectors in the N dimensional Euclidian space.

By definition $M(\mathbf{s})$ is a unit vector whose ith coordinate is 1, where $s^i = \min\limits_{1\leqslant j\leqslant N} s^j$. If there are more than one such i put $M(\mathbf{s})$ to be equal to the minimal of them. By definition $(M(\mathbf{s})\cdot\mathbf{s}) = \min\limits_{1\leqslant j\leqslant N} s^j$.

By definition of the IFPL choice of a leader:

$$r_{1:T} = E\left(\sum_{t=1}^{T} M(\tilde{s}_{1:t})s_t\right).$$

So, we must estimate the sum under the expectation.

We first show that:

$$\sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t}) \cdot \tilde{\mathbf{s}}_t \leqslant M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T}. \tag{4.22}$$

A formal proof is by induction by T. For $T = 1$ this is obvious. For the induction step from $T-1$ to T, we use the following two observations.

We have $\tilde{\mathbf{s}}_{1:T} = \tilde{\mathbf{s}}_{1:T-1} + \tilde{\mathbf{s}}_T$ by definition and

$$M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T-1} \geqslant M(\tilde{\mathbf{s}}_{1:T-1}) \cdot \tilde{\mathbf{s}}_{1:T-1},$$

since the right-hand side of this inequality is equal to the minimal coordinate of the vector $\tilde{\mathbf{s}}_{1:T-1}$, whereas the left-hand side is equal to a coordinate chosen by a different criterion. Combining these observations and the induction hypothesis for step $T-1$, we obtain the induction hypothesis (4.22) for the step T:

$$\begin{aligned} M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T} = M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T-1} + M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_T \geqslant \\ \geqslant M(\tilde{\mathbf{s}}_{1:T-1}) \cdot \tilde{\mathbf{s}}_{1:T-1} + M(\tilde{\mathbf{s}}_{1:T-1}) \cdot \tilde{\mathbf{s}}_T \geqslant \\ \geqslant \sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t}) \cdot \tilde{\mathbf{s}}_t. \end{aligned}$$

Recalling the definition (4.20) of $\tilde{\mathbf{s}}_t$, we rewrite (4.22) as follows:

$$\sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t}) \cdot \mathbf{s}_t \leqslant M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T} + \sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t}) \cdot \xi \left(\frac{1}{\epsilon_t} - \frac{1}{\epsilon_{t-1}} \right) \tag{4.23}$$

Similarly, using the definition (4.21) of $\tilde{\mathbf{s}}_{1:t}$ and a change of criterion for coordinate selection, we obtain

$$\begin{aligned} M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T} \leqslant M(\mathbf{s}_{1:T}) \cdot \left(\mathbf{s}_{1:T} - \frac{\xi}{\epsilon_T} \right) = \\ = \min_{\mathbf{d} \in D} \{ \mathbf{d} \cdot \mathbf{s}_{1:T} \} - \frac{M(\mathbf{s}_{1:T}) \cdot \xi}{\epsilon_T}. \end{aligned} \tag{4.24}$$

By definition $(M(\mathbf{s}_{1:T}) \cdot \xi) = \xi^k$ for some k.

Since $E(\xi) = 1$ for any exponentially distributed variable ξ, the expectation of the subtracted term in (4.24) is equal to

$$E\left(\frac{M(\mathbf{s}_{1:T})\cdot\xi}{\epsilon_T}\right) = \frac{1}{\epsilon_T}E(\xi^k) = \frac{1}{\epsilon_T}. \tag{4.25}$$

The second term of (4.23) satisfies

$$\sum_{t=1}^{T}(M(\tilde{\mathbf{s}}_{\mathbf{1:t}})\cdot\xi)\left(\frac{1}{\epsilon_t}-\frac{1}{\epsilon_{t-1}}\right)\leqslant$$

$$\leqslant\sum_{t=1}^{T}\max_{1\leqslant i\leqslant N}\xi^i\left(\frac{1}{\epsilon_t}-\frac{1}{\epsilon_{t-1}}\right)=\frac{1}{\epsilon_T}\max_{1\leqslant i\leqslant N}\xi^i. \tag{4.26}$$

Here we have used the property $\epsilon_t < \epsilon_{t-1}$ for all t.

We will use the bound for the mathematical expectation E of the maximum of the exponentially distributed variables:

$$0\leqslant E(\max_{1\leqslant i\leqslant N}\xi^i)\leqslant 1+\ln N. \tag{4.27}$$

Indeed, for the exponentially distributed random variables ξ^i, $i = 1,\dots N$,

$$P\{\max_i \xi^i\geqslant a\} = P\{\exists i(\xi^i\geqslant a)\}\leqslant$$

$$\leqslant\sum_{i=1}^{N}P\{\xi^i\geqslant a\} = N\exp\{-a\}. \tag{4.28}$$

The following equality holds for any non-negative random variable η:

$$E(\eta) = \int_0^\infty P\{\eta\geqslant y\}dy. \tag{4.29}$$

For the proof see the problem in Section 4.8.

Then by (4.28) we have

$$E(\max_i \xi^i - \ln N) =$$
$$= \int_0^\infty P\{\max_i \xi^i - \ln N \geqslant y\} dy \leqslant$$
$$\leqslant \int_0^\infty N \exp\{-y - \ln N\} dy = 1.$$

Therefore, $E(\max_i \xi^i) \leqslant 1 + \ln N$. By (4.27) the expectation of (4.26) has the upper bound $\frac{1}{\epsilon_T}(1 + \ln N)$.

Combining the bounds (4.23)–(4.26) and (4.25), we obtain

$$r_{1:T} = E\left(\sum_{t=1}^T M(\tilde{\mathbf{s}}_{\mathbf{1:t}}) \cdot \mathbf{s_t}\right) \leqslant$$
$$\leqslant \min_i s^i_{1:T} + \frac{\ln N}{\epsilon_T}. \tag{4.30}$$

Lemma is proved. $\triangle$.

We finish now the proof of the theorem.

The inequality (4.16) of Lemma 4.3 and the inequality (4.19) of Lemma 4.4 imply the inequality

$$E(s_{1:T}) \leqslant \min_i s^i_{1:T} + a \sum_{t=1}^T \frac{1}{\sqrt{t}} + \frac{1}{a} \ln N\sqrt{T} \leqslant$$
$$\leqslant \min_i s^i_{1:T} + 2a\sqrt{T} + \frac{1}{a} \ln N\sqrt{T}. \tag{4.31}$$

for all T. Optimizing the sum in (4.31) by a, we obtain $a = \sqrt{2 \ln N}$. Hence, we obtain the bound (4.14)

$$E(s_{1:T}) \leqslant \min_i s^i_{1:T} + 2\sqrt{2T \ln N}.$$

The inequality (4.15) directly follows from (4.14). Theorem is proved. $\triangle$

We also obtain some important corollary of this theorem. In this corollary, using some variants of Hoeffding inequality, we will transfer the bound for the expectation of cumulative loss to bound of the cumulative loss itself that holds with probability close to 1.

To do this, we need to complicate the randomization used in the FPL algorithm.

Recall that we have used at each step one and the same sequence of independent and identically distributed random variables $\xi_t^1,\ \ldots,\ \xi_t^N$. We modify randomization in the FPL and IFPL algorithms as follows. Consider an infinite sequence of series of independent identically distributed (according to the exponential law) random variables $\xi_1^t,\ \ldots,\ \xi_N^t$, $t = 1, 2,\ \ldots$, such that all these variables be taken together are independent.

In the algorithm FPL (see Fig. 4.1), we randomize each expert at step t using the series of random variables $\xi_t^1,\ \ldots,\ \xi_t^N$. *Learner* selects the expert which has the minimal perturbed cumulative loss after $t-1$ steps:

$$I_t = \operatorname{argmin}_{i=1,2,\ldots N} \left\{ s^i_{1:t-1} - \frac{1}{\epsilon_t}\xi_t^i \right\}.$$

A similar modification is made in the algorithm IFPL.

In this case, the one-step losses s_t, $t = 1, 2,\ \ldots$, of the FPL algorithm are independent random variables. Proof of Lemma 4.3 remains the same, the proof of Lemma 4.4 changes insignificantly – you just apply the expectation to both parts of inequalities (4.23), (4.24) and (4.26) and use the facts that $E(\xi_t^i) = 1$ and $E(\max_i \xi_t^i) \leqslant 1 + \ln N$ for all i and t.

Corollary 4.1. *Given N and T, for any $\delta > 0$, with probability $1-\delta$,*

$$s_{1:T} \leqslant \min_i s^i_{1:T} + 2\sqrt{2T \ln N} + \sqrt{\frac{T}{2} \ln \frac{1}{\delta}}. \tag{4.32}$$

The FPL algorithm is asymptotically consistent:

$$\limsup_{T\to\infty} \frac{1}{T}(s_{1:T} - \min_{i=1,\ldots N} s^i_{1:T}) \leqslant 0 \tag{4.33}$$

with probability 1.

Proof. To prove the first assertion we use a version of Chernoff inequality given in Corollary 8.3:

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables such that $0 \leqslant X_i \leqslant 1$ for all $i = 1, 2, \ldots$ Then for any $\epsilon > 0$

$$P\left\{\sum_{i=1}^{T} X_i - E\sum_{i=1}^{T} X_i > \epsilon\right\} \leqslant \exp\left(-\frac{2\epsilon^2}{T}\right). \tag{4.34}$$

Put $\delta = \exp\left(-\frac{2\epsilon^2}{T}\right)$. Then $\epsilon = \sqrt{\frac{T}{2}\ln\frac{1}{\delta}}$. For $X_t = s_t$ and by (4.34) we have, with probability $1 - \delta$,

$$\sum_{t=1}^{T} s_t \leqslant E(s_{1:T}) + \sqrt{\frac{T}{2}\ln\frac{1}{\delta}}$$

From this inequality and from the bound (4.14) of Theorem 4.5, we obtain the inequality (4.32).

For the proof of (4.33) we use the Borel–Cantelli lemma and a version of Chernoff inequality (see Section 8.6):

$$P\left\{\left|\frac{1}{T}\sum_{i=1}^{T}(X_i - E(X_i))\right| > \epsilon\right\} \leqslant 2\exp\left(-2T\epsilon^2\right). \tag{4.35}$$

Here we get $X_t = s_t$. Since for any $\epsilon > 0$ the series of exponents from the right hand-part of this inequality converges, by Borel–Cantelli lemma:

$$\limsup_{T\to\infty} \frac{1}{T}(s_{1:T} - E(s_{1:T})) \leqslant 0$$

with probability 1. From this limit and from the bound (4.14) of Theorem 4.5, we obtain the needed inequality (4.15).

Corollary 4.1 is also valid in a more general case of "adversatively adaptive" experts whose losses depend of past values of random variables $s_{t'}$ for $t' < t$. In this case, random variables $X_t = s_t$ are not independent, but $X_t - E(X_t) = s_t - E(s_t)$ form martingale-differences, and we can apply the corresponding Azuma–Hoeffding inequality (8.34) and the strong martingale law of large numbers (8.35).

4.4. Exponentially weighted average forecaster

Let Ω be an outcome set, Γ be a decision set (a prediction set), and Θ be a set experts (experts strategies). Assume that Θ is a finite set, and $\Gamma \subseteq \mathcal{R}^n$. In this section, Ω is an arbitrary set.

Loss from a prediction $\gamma \in \Gamma$ at an outcome $\omega \in \Omega$ is measured by a loss function $\lambda(\omega, \gamma)$ taking non-negative real values. In what follows, we assume that the values of the loss function are in $[0, 1]$.

Consider the perfect-information protocol of the game with players: *Learner*, *Experts*, and *Nature*.
FOR $t = 1, 2, \ \ldots$
Experts θ announce predictions: ξ_t^θ for $\theta \in \Theta$.
Learner announces his prediction: $\gamma_t \in \Gamma$.
Nature announces an outcome: ω_t.
Experts θ update their cumulative losses at step t:

$$L_t(\theta) = L_{t-1}(\theta) + \lambda(\omega_t, \xi_t^\theta)$$

for $\theta \in \Theta$.
Learner updates his cumulative loss at step t:

$$L_t = L_{t-1} + \lambda(\omega_t, \gamma_t).$$

Assume $L_0(\theta) = L_0 = 0$ for all θ.
ENDFOR

This protocol defines the ordering of the players moves. Each player is allowed to determine its action to use all the information known to the beginning of its move.

The *Learner's* goal is to choose a sequence of forecasts $\gamma_1, \gamma_2, \ \ldots$ such that for each t its cumulative loss L_t would be with some degree of accuracy no more than the cumulative loss of the most efficient expert, ie, no more than $\inf\limits_\theta L_t(\theta)$.

Nature can be *adversarial* for *Learner*: her outcomes ω_t can depend on *Learner* forecasts γ_t.

The cumulative regret is defined as

$$R_{\theta,T} = \sup_\theta \sum_{t=1}^{T} (\lambda(\omega_t, \gamma_t) - \lambda(\omega_t, \xi_t^\theta)) = L_T - \inf_\theta L_T(\theta). \qquad (4.36)$$

We call the *Learner's* method of forecasting *asymptotically consistent* if

$$\limsup_{T\to\infty} \frac{1}{T}(L_T - \inf_{\theta} L_T(\theta)) \leqslant 0 \tag{4.37}$$

regardless of the actions of *Nature* and *Experts* i.

Note that *Learner* can predict better even than the expert with the smallest loss.

Assume that forecasts are vectors from n-dimensional Euclidian space $\mathcal{R}^n$. Thus, they can be added and multiplied by real numbers.

A subset Z of the Euclidian space $\mathcal{R}^n$) is called *convex* if for each vectors $z, z' \in Z$ and any real number $0 \leqslant p \leqslant 1$ it holds $pz + (1-p)z' \in Z$.

A function $h(z)$ defined on a convex subset Z is called convex if the set $\{(x, y) : y \geqslant h(x)\}$ is convex. Equivalently, for each $z, z' \in Z$ and for any $0 \leqslant p \leqslant 1$,

$$h(pz + (1-p)z') \leqslant ph(z) + (1-p)h(z'). \tag{4.38}$$

Assume that the decision set Γ is a convex subset of $\mathcal{R}^n$, and the loss function $\lambda(\omega, \gamma)$ is convex by forecast γ.

Assume also, that the set of experts if finite: $\Theta = \{1, \ \ldots, \ N\}$.

The exponential weighted average forecaster outputs a forecast

$$\gamma_t = \frac{\sum\limits_{i=1}^{N} w_{i,t-1}\xi_t^i}{\sum\limits_{j=1}^{N} w_{j,t-1}} = \sum_{i=1}^{N} w_{i,t-1}^* \xi_t^i, \tag{4.39}$$

where $\xi_t^i \in \mathcal{R}^n$ is the forecast of the *Expert* i at step t, $w_{i,t-1}$, $i = 1, \ \ldots, \ N$, are experts weights at step t,

$$w_{i,t-1}^* = \frac{w_{i,t-1}}{\sum\limits_{j=1}^{N} w_{j,t-1}} \tag{4.40}$$

are normalized weights. Since Γ is convex, $\gamma_t \in \Gamma$ for all t.

The weight of experts are defined as

$$w_{i,t-1} = e^{-\eta L_{t-1}(i)}, \tag{4.41}$$

$i = 1, \ \ldots, \ N$, where $L_{t-1}(i)$ is the *Expert* i cumulative loss at steps $\leqslant t-1$, $\eta > 0$ is a parameter or *learning rate*.

Hence, the *Learner* forecast is as follows:

$$\gamma_t = \frac{\sum\limits_{i=1}^{N} \xi_t^i e^{-\eta L_{t-1}(i)}}{\sum\limits_{j=1}^{N} e^{-\eta L_{t-1}(j)}} = \sum_{i=1}^{N} w_{i,t-1}^{*} \xi_t^i, \tag{4.42}$$

where

$$w_{i,t-1}^{*} = \frac{e^{-\eta L_{t-1}(i)}}{\sum\limits_{j=1}^{N} e^{-\eta L_{t-1}(j)}} \tag{4.43}$$

is the weight of *Expert* i, $i = 1, \ \ldots, \ N$.

A performance bound of exponentially weighted average forecaster is given in the following theorem.

Theorem 4.6. *Let $\lambda(\omega, \gamma)$ be a loss function convex in γ with range in $[0, 1]$. Then for any $\eta > 0$, T, and for any sequence of outcomes $\omega_1, \ \ldots, \ \omega_T \in \Omega$, the ciumulative regret of the exponentially weighted average forecaster satisfies the inequality:*

$$L_T - \min_{i=1,\ldots,\ N} L_T^i \leqslant \frac{\ln N}{\eta} + \frac{T\eta}{8}. \tag{4.44}$$

For $\eta = \sqrt{8 \ln N / T}$ the upper bound is $\sqrt{\frac{1}{2} T \ln N}$.

Proof. Define an auxiliary quantity

$$W_t = \sum_{i=1}^{N} w_{i,t} = \sum_{i=1}^{N} e^{-\eta L_t^i}, \tag{4.45}$$

where $W_0 = N$.

We use in this proof the Hoeffding inequality that will be proved in Section 8.6 (see Lemma 8.1 below). This lemma says that for any random variable X such that $a \leqslant X \leqslant b$ and for any real number $s > 0$,

$$\ln E(e^{sX}) \leqslant sE(X) + \frac{s^2(b-a)^2}{8},$$

where E is a synbol of the mathematical expectation.

The proof is based on a comparison of the lower and upper bounds of the quantity $\ln \frac{W_T}{W_0}$.

The lower bound is obtained as follows. Since $w_{i,0} = 1$ for all $i = 1, \ \ldots, \ N$,

$$\ln \frac{W_T}{W_0} = \ln \left(\sum_{i=1}^{N} e^{-\eta L_T^i} \right) - \ln N \geqslant$$

$$\geqslant \ln \left(\max_{i=1, \ ..., \ N} e^{-\eta L_T^i} \right) - \ln N =$$

$$= -\eta \min_{i=1, \ ..., \ N} L_T^i - \ln N. \tag{4.46}$$

The upper bound of the quantity $\ln \frac{W_T}{W_0}$ is obtained as follows. We have for any t,

$$\ln \frac{W_t}{W_{t-1}} = \ln \frac{\sum\limits_{i=1}^{N} e^{-\eta\lambda(\omega_t, \xi_t^i)} e^{-\eta L_{t-1}(i)}}{\sum\limits_{i=1}^{N} e^{-\eta L_{t-1}(i)}} =$$

$$= \ln \frac{\sum\limits_{i=1}^{N} w_{i,t-1} e^{-\eta\lambda(\omega_t, \xi_t^i)}}{\sum\limits_{j=1}^{N} w_{j,t-1}} = \ln E(e^{-\eta\lambda(\omega_t, \xi_t^i)}), \tag{4.47}$$

where E is a mathematical expectation with respect to a probability distribution:

$$w_{i,t-1}^* = \frac{w_{i,t-1}}{\sum\limits_{j=1}^{N} w_{j,t-1}}$$

for $i = 1, \ \ldots, \ N$.

We use the Hoeffding inequality (8.30), where $a = 0$, $b = 1$, and the corresponding random variable X takes the value $\lambda(\omega_t, \xi_t^i)$ with probability $w_{i,t-1}^*$. We use also the convexity property of the loss function $\lambda(\omega, \gamma)$ in the second argument γ. Therefore, by (4.47) we obtain: the following inequalities:

$$\ln \frac{W_t}{W_{t-1}} \leqslant -\eta \frac{\sum\limits_{i=1}^{N} w_{i,t-1} \lambda(\omega_t, \xi_t^i)}{\sum\limits_{j=1}^{N} w_{j,t-1}} + \frac{\eta^2}{8} \leqslant$$

$$\leqslant -\eta \lambda \left(\omega_t, \frac{\sum\limits_{i=1}^{N} w_{i,t-1} \xi_t^i}{\sum\limits_{j=1}^{N} w_{j,t-1}} \right) + \frac{\eta^2}{8} =$$

$$= -\eta \lambda(\omega_t, \gamma_t) + \frac{\eta^2}{8}, \tag{4.48}$$

where γ_t is a forecast of the exponentially weighted average forecaster (4.42).

Summing (4.48) over $t = 1, \ \ldots, \ T$, we obtain:

$$\ln \frac{W_T}{W_0} = \sum_{i=1}^{T} \ln \frac{W_t}{W_{t-1}} \leqslant -\eta L_T + \frac{\eta^2}{8} T. \tag{4.49}$$

Using the lower (4.46) and the upper (4.49) bounds, we obtain

$$L_T \leqslant \min_{i=1,\ldots,\ N} L_T^i + \frac{\ln N}{\eta} + \frac{\eta}{8} T. \tag{4.50}$$

theorem is proved. $\triangle$

Note that for $\eta = \sqrt{8 \ln N / T}$, the cumulative regret is bounded by $\sqrt{\frac{1}{2} T \ln N}$.

The obvious drawback of this bound is that the prediction horizon T is used for defining the parameter η.

A uniform bound, based on the use of variable learning rate, is given in the next section.

4.5. Exponentially weighted average forecaster with variable learning rate

In this section we consider a technically more complicated algorithm of the exponentially weighted average forecaster with a variable learning rate. This construction was proposed by Alexey Chernov [8].

Recall that L_T^i is the cumulative loss of the ith expert at the first T steps, L_T is the cumulative loss of *Learner*, $\Theta = \{1,\ \ldots,\ N\}$ – the set of all experts, Γ is a convex prediction set that is a subset of $\mathcal{R}^n$, $\lambda(\omega, \gamma)$ is a loss function convex on γ.

We modify the exponentially weighted average forecaster. Now, weights are defined as

$$w_{i,t-1} = e^{-\eta_t L_{t-1}^i},$$

$i = 1,\ \ldots,\ N$, where L_{t-1}^i is the cumulative loss of the ith expert at steps $\leqslant t-1$, $\eta_t > 0$ is a variable learning rate.

In this case we can obtain a uniform upper bound of the regret.

Theorem 4.7. *For any T anf for any sequence of positive real numbers $\eta_1 \geqslant \eta_2 \geqslant \ldots, \eta_T$, and for any sequence of outcomes $\omega_1, \ldots, \omega_T \in \Omega$, the regret of the exponentially weighted average forecaster with variable learning rate η_t satisfies*

$$L_T - \min_{i=1,\ldots,N} L_T^i \leqslant \frac{\ln N}{\eta_T} + \frac{1}{8}\sum_{t=1}^{T} \eta_t\,. \tag{4.51}$$

In particular, for $\eta_t = \sqrt{\frac{4\ln N}{t}}$, $t = 1, \ldots, T$, it holds

$$L_T - \min_{i=1,\ldots,N} L_T^i \leqslant \sqrt{T \ln N}\,.$$

Proof. At any step t *Learner* outputs a forecast $\widehat{p}_t = \sum_{i=1}^N \xi_t^i w_{i,t-1}/W_{t-1}$, where $w_{i,t-1} = e^{-\eta_t L_{t-1}(i)}$ and $W_{t-1} = \sum_{j=1}^N w_{j,t-1}$. By convexity of $\lambda(\omega, \gamma)$ in the second argument, we obtain

$$\lambda(\omega_t, \widehat{p}_t) \leqslant \sum_{i=1}^{N} \frac{w_{i,t-1}}{W_{t-1}}\ \lambda(\omega_t, \xi_t^i)\,.$$

By the Hoeffding inequality, we obtain

$$e^{-\eta_t \sum_{i=1}^N \frac{w_{i,t-1}}{W_{t-1}} \lambda(\omega_t, \xi_t^i)} \geqslant \sum_{i=1}^N \frac{w_{i,t-1}}{W_{t-1}} e^{-\eta_t \lambda(\omega_t, \xi_t^i) - \eta_t^2/8}$$

Rewrite this inequality as follows

$$e^{-\eta_t \lambda(\omega_t, \widehat{p}_t)} \geqslant \sum_{i=1}^N \frac{w_{i,t-1}}{W_{t-1}} e^{-\eta_t \lambda(\omega_t, \xi_t^i) - \eta_t^2/8} \,. \tag{4.52}$$

Define the auxiliary quantities:

$$s_{i,t-1} = e^{-\eta_{t-1} L_{t-1}(i) + \eta_{t-1}\left(\widehat{L}_{t-1} - \frac{1}{8}\sum_{k=1}^{t-1} \eta_k\right)}$$

and note that

$$\frac{w_{i,t-1}}{W_{t-1}} = \frac{\frac{1}{N}(s_{i,t-1})^{\frac{\eta_t}{\eta_{t-1}}}}{\sum_{j=1}^N \frac{1}{N}(s_{j,t-1})^{\frac{\eta_t}{\eta_{t-1}}}} \,. \tag{4.53}$$

Let us show that $\sum_{j=1}^N \frac{1}{N} s_{j,t} \leqslant 1$ by mathematical induction over t. For $t = 0$ this is trivial, since $s_{i,0} = 1$ for all i. Assume that $\sum_{j=1}^N \frac{1}{N} s_{j,t-1} \leqslant 1$. Then

$$\sum_{j=1}^N \frac{1}{N}(s_{j,t-1})^{\frac{\eta_t}{\eta_{t-1}}} \leqslant \left(\sum_{j=1}^N \frac{1}{N} s_{j,t-1}\right)^{\frac{\eta_t}{\eta_{t-1}}} \leqslant 1 \,, \tag{4.54}$$

since the function $x \mapsto x^\alpha$ is concave and monotone for $x \geqslant 0$ and $\alpha \in [0,1]$ and since $0 \leqslant \eta_t \leqslant \eta_{t-1}$. Using (4.54) to bound the right-hand side of (4.53), we get $w_{i,t-1}/W_{t-1} \geqslant (s_{i,t-1})^{\frac{\eta_t}{\eta_{t-1}}}/N$; and combining with (4.52), we get

$$e^{-\eta_t \ell(\widehat{p}_t, y_t)} \geqslant \sum_{i=1}^N \frac{1}{N}(s_{i,t-1})^{\frac{\eta_t}{\eta_{t-1}}} e^{-\eta_t \ell(f_{i,t}, y_t) - \eta_t^2/8} \,.$$

It remains to note that

$$s_{i,t} = (s_{i,t-1})^{\frac{\eta_t}{\eta_{t-1}}} e^{-\eta_t \ell(f_{i,t}, y_t) + \eta_t \ell(\widehat{p}_t, y_t) - \eta_t^2/8}$$

and we get $\sum_{j=1}^{N} \frac{1}{N} s_{j,t} \leqslant 1$.

For any i we have $\frac{1}{N} s_{i,n} \leqslant \sum_{j=1}^{N} \frac{1}{N} s_{j,n} \leqslant 1$, thus

$$-\eta_T L_{i,T} + \eta_T \left(\widehat{L}_T - \frac{1}{8} \sum_{t=1}^{T} \eta_t \right) \leqslant \ln N$$

and (4.51) follows. $\triangle$

A problem from Section 4.8 asserts that the bound $L_T - \min_{i=1,\ldots,N} L_T^i \leqslant \sqrt{T \ln N}$ of Theorem 4.7 is also valid for the allocation algorithm $Hedge(\beta_t)$ with a variable learning rate $\beta_t = e^{-\eta_t}$, where $\eta_t = \sqrt{\frac{4 \ln N}{t}}$.

4.6. Randomized forecasting

Assume that an outcome set Ω and a loss function $\lambda(\omega, \gamma)$ be given. The results of this section can be applied to loss functions not convex by the forecast.

Assume that there are N experts. Recall the deterministic forecasting protocol of prediction with expert advice.
Let $L_0 = 0$, $L_0(i) = 0$, $i = 1, \ \ldots, \ N$.
FOR $t = 1, 2, \ldots$
Expert i announces a forecast $\xi_t^i \in \Gamma$, $i = 1, \ldots, \ N$.
Learner announces a forecast $\gamma_t \in \Gamma$.
Nature announces an outcome $\omega_t \in \Omega$.
Expert i updates its cumulative loss at step t:

$$L_t(i) = L_{t-1}(i) + \lambda(\omega_t, \xi_t^i),$$

where $i = 1, \ldots, \ N$.
Learner updates its cumulative loss at step t:

$$L_t = L_{t-1} + \lambda(\omega_t, \gamma_t).$$

ENDFOR

This is a perfect information protocol: each player can use all the information known at the beginning of his move.

The *Learner* cumulative loss at steps $t = 1, \ \ldots, \ T$ is equal to

$$L_T = \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t).$$

The *Expert* i cumulative loss at steps $t = 1, \ \ldots, \ T$ is equal to

$$L_T(i) = \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i).$$

Example. We give an example which shows that for some games with non-convex loss functions $\lambda(\omega, \gamma)$ each method of deterministic forecasting has an unacceptably high cumulative regret which grows linearly with the length of the period of prediction.

Consider a simple game with two experts 1 and 2. Outcomes and prediction spaces are the same: $\Omega = \Gamma = \{1, 2\}$. The loss fuunction is $\lambda(\omega, \gamma) = 1_{\{\omega \neq \gamma\}}$ – the characteristic function of the set $\{(\omega, \gamma) : \gamma \neq \omega\}$. Clearly, this loss function is trivially not convex.

Note that for any deterministic strategy of *Learner* $\gamma_1, \gamma_2, \ \ldots$ a sequence of outcomes $\omega_1, \omega_2, \ \ldots$ exists such that *Learner* suffers the maximal possible loss: $L_T = T$ for all T. Indeed, *Nature* can define

$$\omega_t = \begin{cases} 2 \text{ if } \gamma_t = 1, \\ 1 \text{ otherwise} \end{cases}$$

for all $t = 1, 2, \ \ldots$.

Consider two experts, one of which - *Expert 1*, always predicts $\xi_t^1 = 1$ and the other - *Expert 2*, always predicts $\xi_t^2 = 2$, $t = 1, 2, \ \ldots$.

Let $L_t(i)$ be the total loss of the ith expert, $i = 1, 2$.

Note that *Learner* just follows the decision of the *Expert 1*, when $\gamma_t = 1$, and follows the decision of *Expert 2* if $\gamma_t = 2$.

It is easy to see that for any sequence of outcomes $\omega_1, \omega_2, \ \ldots, \ \omega_t$ number of ones or number of twos will be more than $t/2$. Then one of these two experts suffers loss no more than $t/2$, and so, $\min_{i=1,2} L_t(i) \leqslant t/2$ for all t.

Therefore, for any sequence of forecasts issued by *Learner* the "adversatively adaptive" *Nature* can output a sequence of outcomes

$\omega_1, \omega_2, \ \ldots$ such that

$$L_T - \min_{i=1,2} L_t(i) \geqslant T/2$$

for all T.

This example shows that for some non-convex loss functions *Nature* can produce a sequence of outcomes such that for any sequence of deterministic predictions the regret is $\geqslant T/2$ for any period T.

This drawback can be overcome by randomization of the *Learner* forecasts. More precisely, the forecasts will be mixed strategies – probability distributions on the set of all deterministic forecasts. We replace the loss function on its expectation and apply results of Section 4.2.

Now suppose that at each step t of the game the *Learner's* forecast is a mixed strategy – a probability distribution $\bar{p}_t = \{p_{1,t}, \ \ldots, \ p_{N,t}\}$ on the set $\{1, \ \ldots, \ N\}$ of experts.

We introduce one more player – *Random Number Generator* that will generate the elements of the set $\{1, \ \ldots, \ N\}$ according to a given probability distribution.

Protocol of the randomized game is as follows.
Define $L_0 = 0$, $L_0(i) = 0$, $i = 1, \ \ldots, \ N$.
FOR $t = 1, 2, \ \ldots$
Expert i announces a forecast $\xi_t^i \in \Gamma$, $i = 1, \ \ldots, \ N$.
Learner announces a probability distribution $\bar{p}_t = \{p_{1,t}, \ \ldots, \ p_{N,t}\}$ on the set of all experts $\{1, \ \ldots, \ N\}$.
Nature announces an outcome $\omega_t \in \Omega$.
Random Number Generator announces an expert $i_t \in \{1, \ \ldots, \ N\}$ with probability $p_{i,t}$.
Expert i updates its cumulative loss at step t:

$$L_t(i) = L_{t-1}(i) + \lambda(\omega_t, \xi_t^i),$$

for $i = 1, \ \ldots, \ N$.
Learner updates its cumulative loss at step t:

$$L_t = L_{t-1} + \lambda(\omega_t, \xi_t^{i_t}).$$

ENDFOR

We can define a random number generator with a given probability distribution using the random number generator producing uniformly distributed real numbers from the unit interval $[0,1]$ as follows.

Define the random variables I_t such that $I_t = i$ if and only if

$$U_t \in \left[\sum_{j=1}^{i-1} p_{j,t}, \sum_{j=1}^{i} p_{j,t} \right],$$

where $U_1, U_2, \ldots$ are independent uniformly distributed in $[0,1]$ random variables. By definition $P\{I_t = i\} = p_{i,t}$ for all t.

In such a game the *Learner's* loss $\lambda(\omega_t, \xi_t^{I_t})$ is a random variable. In this case, the *Learner's* performance is evaluated by a random variable – random regret, as follows:

$$L_T - \min_{i=1,\ldots,N} L_T(i) = \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^{I_t}) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i). \quad (4.55)$$

Consider the setting in which *Learner* minimizes the mathematical expectation of the regret (4.55):

$$\begin{aligned} E(L_T - \min_{i=1,\ldots,N} L_T(i)) = \\ = E(L_T) - \min_{i=1,\ldots,N} L_T(i) = \\ = \sum_{t=1}^{T} E(\lambda(\omega_t, \xi_t^{I_t})) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i) = \\ = \sum_{t=1}^{T} \sum_{i=1}^{N} \lambda(\omega_t, \xi_t^i) p_{i,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i). \end{aligned} \quad (4.56)$$

We shall calculate the probability distribution on the set of experts with the help of the definition (4.4) from Section 4.2. At step t, define

$$p_{i,t} = \frac{\beta^{\sum_{s=1}^{t-1} l_s^i}}{\sum_{j=1}^{N} \beta^{\sum_{s=1}^{t-1} l_s^j}}, \quad (4.57)$$

where $l_s^i = \lambda(\omega_s, \xi_s^i)$ for $i = 1, \ldots, N$, $0 < \beta < 1$.

The algorithm (4.57) is called *randomized exponentially weighted average forecaster* From Lemma 4.2, we obtain:

Theorem 4.8. *Let L_T be a random variable representing the cumulative loss of the randomized version of the algorithm $Hedge(\beta)$ at T steps, where $\beta = g(T/\ln N)$ for*

$$g(x) = \frac{1}{1 + \sqrt{\frac{2}{x}}}.$$

Then the mathematical expectation of the cumulative loss of the randomized exponentially weighted average forecaster is bounded by

$$E(L_T) \leqslant \min_i L_T(i) + \sqrt{2T \ln N} + \ln N. \tag{4.58}$$

The drawback of this bound is that the parameter β depends on the horizon T. See also the comment at the end of Section 4.4.

Note that for each t the probability vector $\bar{p}_t = \{p_{1,t}, \ldots, p_{N,t}\}$ depends on the sequence of previous outcomes $\omega_1, \ldots, \omega_{t-1}$ issued by *Nature*, and the sequence $\omega_1, \ldots, \omega_t$, in turn, may depend on the sequence of distributions $\bar{p}_s = \{p_{1,s}, \ldots, p_{N,s}\}$, $s = 1, \ldots t$ issued by *Learner*.

By Ionesco–Tulcea theorem (see [29]) there is an overall probability distribution $\mathcal{P}$ defined on infinite paths of experts $i_1, i_2, \ldots$, where $i_t \in \{1, \ldots N\}$ for all $t = 1, 2, \ldots$, generated by probability distributions $\bar{p}_t = \{p_{1,t}, \ldots, p_{N,t}\}$, $t = 1, 2, \ldots$.

Using Corollary 8.5 from the Azuma–Hoeffding inequality (see Lemma 8.2 below) and the inequality (4.58) we can obtain the following corollary.

Corollary 4.2. *For any $0 < \delta < 1$ and T, with probability $1 - \delta$, the regret of the randomized exponentially weighted average forecaster satifies the inequality:*

$$\sum_{t=1}^{T} \lambda(\omega_t, \xi_t^{I_t}) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i) \leqslant$$

$$\leqslant \sqrt{2T \ln N} + \ln N + \sqrt{\frac{1}{2} T \ln \frac{1}{\delta}}.$$

Proof. By definition the sequence of random variables

$$X_t = \lambda(\omega_t, \xi_t^{I_t}) - E(\lambda(\omega_t, \xi_t^{I_t})) =$$
$$= \lambda(\omega_t, \xi_t^{I_t}) - \sum_{i=1}^{N} \lambda(\omega_t, \xi_t^i) p_{i,t}$$

is a sequence of the bounded martingale-differences. By Corollary 8.5 their sums

$$S_T = \sum_{t=1}^{T} X_t$$

satisfy the inequality

$$P\{S_T > c\} \leqslant e^{-\frac{2c^2}{T}} \tag{4.59}$$

for all T, where c is a positive real number (see (8.33)). Then for any $\delta > 0$ the following inequality holds:

$$P\left\{S_T > \sqrt{\frac{1}{2} T \ln \frac{1}{\delta}}\right\} \leqslant \delta.$$

The corollary now follows directly from this inequality and the inequalities (4.56) and (4.58). $\triangle$

Let in some game of prediction with experts *Expert* i, $i = 1, \ldots, N$, outputs the forecasts $\xi_1^i, \xi_2^i, \ldots$, and *Learner* outputs the forecasts $\xi_1, \xi_2, \ldots$.

A randomized forecasting algorithm is called Hannan consistent if, with $\mathcal{P}$-probability one:

$$\limsup_{T\to\infty} \frac{1}{T} \left(\sum_{t=1}^{T} \lambda(\omega_t, \xi_t) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i) \right) \leqslant 0. \tag{4.60}$$

The following corollary is proved just as well as the statement (4.15), where a version of the algorithm $Hedge(\beta_t)$ with a variable learning rate is used (see a problem in Section 4.8).

Corollary 4.3. *The randomized exponentially weighted average forecaster with a variable learning rate is Hannan consistent.*

Let us explain the relation between the example given at the beginning of this section and Corollary 4.3.

In the example, if $\gamma_t = 1$ then *Learner* simply follows the expert $i_t = 1$ prediction, and if $\gamma_t = 2$ then *Learner* follows the expert $i_t = 2$ prediction. Therefore, we obtain an infinite trajectory of chosen experts: $i_1, i_2, \ldots$. When experts are chosen at random, the $\mathcal{P}$-probability to choose this trajectory, as well as any other, which violates the condition (4.60) is 0.

A comparison with Theorem 4.2 shows that the randomized algorithm when be applied to a simple loss function, has about twice smaller upper bound of regret than the deterministic weighted majority algorithm WMA (see problems in Section 4.8).

4.7. Boosting

This section describes method of reinforcement of simple classifiers, called *boosting*. This method is based on combining primitive *weak classifiers* into a single *strong classifier*. Under the strength of a classifier we mean the average number of classification errors made on a training set.

We combine weak classifiers using the method of prediction with expert advice.

A weak learner is defined to be a classifier which is only slightly correlated with the true classification (it can label examples better than random guessing). In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification.

We will study the algorithm AdaBoost proposed by Freund and Shapire [14]. It is a meta-algorithm, and can be used in conjunction with many other learning algorithms to improve their performance. The classifiers it uses can be weak (ie, display a substantial error rate), but as long as their performance is not random (resulting in an error rate of 0.5 for binary classification), they will improve the final model.

AdaBoost is adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost generates and calls a new weak classifier in each of

a series of rounds. For each call, a distribution of weights is updated that indicates the importance of examples in the data set for the classification. On each round, the weights of each incorrectly classified example are increased, and the weights of each correctly classified example are decreased, so the new classifier focuses on the examples which have so far eluded correct classification.

This method for amplification of simple classifiers used in many applications, and is still the subject of many both applied and theoretical research.

4.7.1. AdaBoost

In this section, we present and analyze a boosting algorithm inspired by the methods we used in Section 4.2 for solving the online allocation problem.

Formally, boosting proceeds as follows. The booster is provided with a set of labeled training examples $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$, where $x_i \in \mathcal{X}$ and $y_i \in Y$. We suppose that $Y = \{0, 1\}$, and a structure of probability space is defined on the set $\mathcal{X} \times Y$. We also suppose that the pairs (x_i, y_i) are i.i.d. according to some fixed but unknown to us probability distribution P. As usual, the goal is to learn to predict the label y given an instance x.

A strong learning algorithm is an algorithm that, given $\epsilon, \delta > 0$ and access to random sample S, outputs with probability $1 - \delta$ a classification hypothesis h_S with error at most ϵ. Further, the running time must be polynomial in $1/\epsilon$, $1/\delta$ and sample size l.

A weak learning algorithm satisfies the same conditions but only for $\epsilon \leqslant \frac{1}{2} - \gamma$ where $\gamma > 0$ is either a constant, or decreases as $1/p$ where p is a polynomial in the relevant parameters. We use WeakLearn to denote a generic weak learning algorithm.

Here we consider only the problem of constructing a classification hypothesis h_S by the training set S. The problem of evaluation of its predictive performance will not be discussed in this section. We refer reader to Freund and Shapire [14], where some bounds of predictive performance of the boosting algorithm in terms of VC-dimension are given.

Let $D(i)$ be a probability distribution over the training examples

(more correctly, it is a distribution over their indices). By definition $D(i) \geqslant 0$ for all $1 \leqslant i \leqslant l$ and

$$\sum_{i=1}^{l} D(i) = 1.$$

Ordinarily, this distribution will be set to be uniform so that $D(i) = 1/N$ for all i.

The empirical error of a classifier h on a training sample S with respect to a distribution D is defined as

$$\epsilon = D\{i : h(x_i) \neq y_i\} = \sum_{i:h(x_i)\neq y_i} D(i).$$

In particular, for the uniform distribution $D(i) = 1/l$ this empirical error is equal to the portion of mistakes:

$$\epsilon = |\{i : h(x_i) \neq y_i\}|/l.$$

Some learning algorithms can be generalized to use a given distribution D directly. For instance, gradient based algorithms and some implementations of support vector machines can use the probability associated with each example to scale the update step size which is based on the example. If the algorithm cannot be generalized in this way, the training sample can be *resampled* to generate a new set of training examples that are distributed according to the given probability distribution.

The boosting algorithm AdaBoost is described below. The goal of this algorithm is to find a final hypothesis with low error relative to a given distribution D over the training examples. Unlike the distribution P, which is over $\mathcal{X} \times Y$ and is set by "nature", the distribution D is only over the instances in the training set and is controlled by the learner.

Input: a sample $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$, a distribution D over $\{1, \ldots, l\}$, weak learning algorithm WeakLearn, integer T specifying number of iterations.

Initialize: the weight vector: $w_1^i = D(i) \ \ i = 1, \ \ldots, \ l$.

FOR $t = 1, \ \ldots, \ T$
1) **Set** for $i = 1, \ \ldots, \ l$

$$p_t^i = \frac{w_t^i}{\sum_{j=1}^{l} w_t^j}.$$

2) **Call** the algorithm WeakLearn providing it with the distribution $D(i) = p_t^i$ for $1 \leqslant i \leqslant l$, get back a classification hypothesis h_t.
3) **Calculate** the empirical error of h_t :

$$\epsilon_t = \sum_{i=1}^{l} p_t^i |h_t(x_i) - y_i|.$$

4) **Set** $\beta_t = \epsilon_t/(1 - \epsilon_t)$.
5) **Set** the new weights vector to be

$$w_{t+1}^i = w_t^i \beta_t^{1-|h_t(x_i)-y_i|}$$

for $i = 1, \ \ldots, \ l \ :$
ENDFOR

Output the hypothesis:

$$h(x) = \begin{cases} 1 \text{ if } f(x) \geqslant \frac{1}{2}, \\ 0 \text{ otherwise,} \end{cases}$$

where the threshold function f is defined as a linear combination of the outputs of the T weak hypotheses using a weighted majority vote

$$f(x) = \sum_{t=1}^{T} q_t h_t(x),$$

with weights

$$q_t = \frac{\ln(1/\beta_t)}{\sum_{t=1}^{T} \ln(1/\beta_t)},$$

for $t = 1, \ldots, \ T$.

This algorithm is a version of the optimal online allocation algorithm $Hedge(\beta)$ defined in Section 4.2 with a dynamical changing parameter β. There is "a dual" relationship between the online allocation algorithm and the boosting algorithm. Put another way, there is a direct mapping or reduction of the boosting problem to the online allocation problem. In such a reduction, one might naturally expect a correspondence relating the strategies to the weak hypotheses and the trials (and associated loss vectors) to the examples in the training set. However, this reduction is reversed: the "strategies" correspond to the examples, and the trials are associated with the weak hypotheses. Another reversal is in the definition of the loss: in $Hedge(\beta)$ the loss l_t^i is small if the ith strategy suggests a good action on the tth trial while in AdaBoost the "loss" $l_t^i = 1 - |h_t(x_i) - y_i|$ appearing in the weight-update rule is small if the tth hypothesis suggests a bad prediction on the ith example. The reason is that in $Hedge(\beta)$ the weight associated with a strategy is increased if the strategy is successful while in AdaBoost the weight associated with an example is increased if the example is "hard."

Thus, the algorithm AdaBoost detects examples on which the algorithm WeakLearn gives wrong classification, and forces it to learn from these examples.

In the analysis, the following property weak algorithm WeakLearn will be used - for any distribution on the training examples the empirical error is less than $1/2$ up to some positive value γ.

The algorithm AdaBoost is analyzed in the following theorem.

Theorem 4.9. *Suppose the weak learning algorithm* WeakLearn*, when called by* AdaBoost *on steps* $t = 1, \ldots, T$*, generates hypotheses with errors* $\epsilon_1, \ldots, \epsilon_T$ *with respect to corresponding distributions* $\bar{p}_1 = \bar{D}, \bar{p}_2, \ldots, \bar{p}_T$*. Then the empirical error*

$$\epsilon = D\{h(x_i) \neq y_i\} = \sum_{h(x_i) \neq y_i} D(i)$$

of the final hypothesis h *output by* AdaBoost *is bounded above by*

$$\epsilon \leqslant 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t(1-\epsilon_t)}. \tag{4.61}$$

Proof. Just as in the proof of Lemmas 4.1 and 4.2 from Section 4.2 we estimate above and below the value of $\sum\limits_{i=1}^{l} w_{T+1}^i$. We have an upper bound:

$$\sum_{i=1}^{l} w_{t+1}^i = \sum_{i=1}^{l} w_t^i \beta_t^{1-|h_t(x_i)-y_i|} \leqslant$$

$$\leqslant \sum_{i=1}^{l} w_t^i (1-(1-\beta_t)(1-|h_t(x_i)-y_i|)) =$$

$$= \left(\sum_{i=1}^{l} w_t^i \right) (1-(1-\beta_t)(1-\epsilon_t)). \tag{4.62}$$

Using (4.62) T times, we obtain

$$\sum_{i=1}^{l} w_{T+1}^i \leqslant \prod_{t=1}^{T} (1-(1-\beta_t)(1-\epsilon_t)). \tag{4.63}$$

Here we have used the definition of the empirical error ϵ_t of the algorithm WeakLearn at step t:

$$\epsilon_t = \sum_{i=1}^{l} p_t^i |h_t(x_i)-y_i| = \sum_{i=1}^{l} \left(\frac{w_t^i}{\sum\limits_{j=1}^{l} w_t^j} \right) |h_t(x_i)-y_i|.$$

Lemma 4.5. *The resulting classifier h makes a mistake at an object x_i if and only if*

$$\prod_{t=1}^{T} \beta_t^{-|h_t(x_i)-y_i|} \geqslant \left(\prod_{t=1}^{T} \beta_t \right)^{-1/2}. \tag{4.64}$$

Proof. Indeed, this statement follows directly from the definition of the classifier h when $y_i = 0$, since in this case $\beta_t^{-|h_t(x_i)-y_i|} = \beta_t^{-h_t(x_i)}$ for all t.

By definition, the equality $h(x_i) = 1$ is valid if and only if

$$\sum_{t=1}^{T} \ln(1/\beta_t) h_t(x_i) \geqslant \frac{1}{2} \sum_{t=1}^{T} \ln(1/\beta_t). \tag{4.65}$$

The inequality (4.65) is equivalent to the inequality (4.64).

Now let $y_i = 1$. Then $h_t(x_i) \leqslant y_i$ for all t. Thus, $\beta_t^{-|h_t(x_i)-y_i|} = \beta_t^{-(1-h_t(x_i))}$ for all t. In this case

$$\beta_t^{-|h_t(x_i)-y_i|} = \beta_t^{-1+h_t(x_i)}. \tag{4.66}$$

for all $1 \leqslant t \leqslant T$.

By definition the equality $h(x_i) = 0$ can only be valid if

$$\prod_{t=1}^{T} \beta_t^{-h_t(x_i)} < \left(\prod_{t=1}^{T} \beta_t \right)^{-1/2}. \tag{4.67}$$

The inequality (4.67) is equivalent to the inequality

$$\prod_{t=1}^{T} \beta_t^{h_t(x_i)} > \left(\prod_{t=1}^{T} \beta_t \right)^{1/2}. \tag{4.68}$$

The equality (4.66) and the inequality (4.68) imply the inequality (4.64). Lemma is proved. $\triangle$

Returning to the proof of the theorem, we note that by definition

$$w_{T+1}^i = D(i) \prod_{t=1}^{T} \beta_t^{1-|h_t(x_i)-y_i|}. \tag{4.69}$$

By Lemma 4.5, (4.64), and (4.69) we obtain

$$\begin{aligned} \sum_{i=1}^{l} w_{T+1}^i \geqslant \sum_{i:h(x_i)\neq y_i} w_{T+1}^i \geqslant \\ \geqslant \left(\sum_{i:h(x_i)\neq y_i} D(i) \right) \left(\prod_{t=1}^{T} \beta_t \right)^{1/2} = \\ = \epsilon \left(\prod_{t=1}^{T} \beta_t \right)^{1/2}, \end{aligned} \tag{4.70}$$

where ϵ is the empirical error of final classification hypothesis h with respect to D.

Combining (4.63) with (4.70), we obtain

$$\epsilon \leqslant \prod_{t=1}^{T} \frac{1-(1-\beta_t)(1-\epsilon_t)}{\sqrt{\beta_t}}. \tag{4.71}$$

Since all factors of the product (4.71) are non-negative, we can minimize by β_t each factor separately. Equate to zero the first derivative by β_t :

$$\frac{d}{d\beta_t}\left(\frac{1-(1-\beta_t)(1-\epsilon_t)}{\sqrt{\beta_t}}\right) = 0.$$

Solving equation with respect to β_t, we obtain: $\beta_t = \epsilon_t/(1-\epsilon_t)$. Putting this expression into (4.71), we obtain (4.61). Theorem is proved. $\triangle$

Corollary 4.4. *The empirical error of the resulting classifier h satisfies the inequality*

$$\epsilon \leqslant \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right), \tag{4.72}$$

where $\epsilon_t = \frac{1}{2} - \gamma_t$, $\gamma_t > 0$ for $t = 1, \ \ldots, \ T$.

In the case, where $\gamma_t = \gamma$ for all t, the inequality (4.72) *reduces to*

$$\epsilon \leqslant \exp(-2T\gamma^2). \tag{4.73}$$

Proof. Indeed, the bound (4.61) from Theorem 4.9 at $\epsilon_t = \frac{1}{2} - \gamma_t$ becomes

$$2\sqrt{\epsilon_t(1-\epsilon_t)} = \sqrt{1-4\gamma_t^2}.$$

Then

$$\epsilon \leqslant \prod_{t=1}^{T} \sqrt{1-4\gamma_t^2} =$$

$$= \exp\left(\sum_{t=1}^{T} \frac{1}{2}\ln(1-4\gamma_t^2)\right) \leqslant$$

$$\leqslant \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right). \tag{4.74}$$

The inequality (4.72) is proved.

To prove (4.73) note that the inequality (4.72) for $\gamma_t = \gamma$ becomes

$$\epsilon \leqslant (1-4\gamma^2)^{T/2} = \exp((T/2)\ln(1-4\gamma^2)) \leqslant \exp(-2T\gamma^2).$$

The exponential upper bound (4.73) allows us to estimate the number of iterations of the algorithm AdaBoost required to achieve the learning error $\leqslant \epsilon$ of the resulting classifier h:

$$T \geqslant \frac{1}{2\gamma^2}\ln\frac{1}{\epsilon}.$$

4.7.2. Laboratory work

Write a computer program realizing the algorithm AdaBoost using the shelf software for SVM described in Section 2.13 as the weak learning algorithm. Apply it to strengthen the algorithm of recognition of handwritten digits from the website: http://www.cs.toronto.edu

4.8. Problems

1. Construct a variant of the weighted majority algorithm for the case when an expert exists in the pool which makes no more than k mistakes. Compute a performance bound for this algorithm.

2. Consider the protocol of the game of prediction with expert advice, where *Nature* outputs a sequence $0^T(01)^T1^T$. There are three constant experts. *Expert 1* outputs $\xi_t^1 = 0$ for all $t = 1, \ldots, 4T$,

Expert 2 outputs $\xi_t^2 = 1$ for all $t = 1, \ldots, 4T$, *Expert 3* outputs $\xi_t^3 = 1/2$ for all $t = 1, \ldots, 4T$. The loss function is $\lambda(\omega, \gamma) = |\omega - \gamma|$.

Compute at each time points $t = 1, \ldots, 4T$:

(i) the weights of experts and their losses;

(ii) the loss *Allocator* and the prediction and cumulative loss of the exponentially weighted forecaster.

3. Compute the forecasts of the exponential weighted forecaster for the case of quadratic and absolute loss functions, where $\Omega = \{0, 1\}$ is the set of outcomes and $\Gamma = [0, 1]$ is the prediction set.

4. Check the simplest properties of the exponential distribution with density $p(x) = e^{-x}$: $P\{\xi > a\} = e^{-a}$ and $P\{\xi > a + b\} = e^{-b}P\{\xi > a\}$ for all nonnegative a and b.

5. Prove that for any non-negative random variable η with a density $p(t)$ the following equality is valid:

$$E(\eta) = \int_0^\infty P\{\eta \geqslant y\}dy.$$

Note. Use $p(y) = F'(y)$, where $F(y) = \int_0^y p(t)dt = 1 - P\{\eta \geqslant y\}$. After that, apply the integration by parts of $E(\eta) = \int_0^\infty tp(t)dt$.

6. Prove Lemma 4.4 for the case where an infinite sequence $\xi_t^1, \ldots, \xi_t^N$, $t = 1, 2, \ldots$ is used for randomization in the algorithms FPL and IFPL.

7. Formulate and study the randomized versions of the weighted majority algorithm WMA. In particular:

a) use a randomized version of the online allocation algorithm $Hedge(\beta)$ from Section 4.2 and apply it to the simple loss function;

b) use follow the perturbed leader algorithm and apply it to the simple loss function.

Formulate and prove the corresponding versions of Theorem 4.2. Compare the performance of the randomized algorithms with the performance of the deterministic weighted majority algorithm WMA (*Hint:* Compute the expected cumulative losses of randomized algorithms and apply Hoeffding inequality).

8. Show that the online allocation algorithm from Section 4.2 is a special case of the exponential weighted forecaster defined in Section 4.4.

(*Hint:* Consider the outcome set $\Omega = [0,1]^N$ consisting of the vectors $\bar{l} = (l^1, \ldots, l^N)$ of expert losses, where N – is the numbes of experts. The prediction set is the simplex Γ of all probability distributions $\bar{p} = (p^1, \ldots, p^N)$, and the loss function is $\lambda(\bar{l}, \bar{p}) = (\bar{l} \cdot \bar{p})$ that is the dot product of the vectors $\bar{l} \in [0,1]^N$ and $\bar{p} \in \Gamma$.

A forecast of the ith expert is an N-dimensional unit vector $\bar{\xi}^i = (0, \ldots, 1, \ldots, 0)$. Then the forecast of of the exponential weighted forecaster at any step t, that was defined in Section 4.4, can be represented in the form:

$$\bar{\xi}_t = \sum_{i=1}^{N} \bar{\xi}_t^i w_{i,t}^* = \bar{p}_t,$$

where $\bar{p}_t = (w_{1,t}^*, \ldots, w_{N,t}^*)$ is a vector of normalized weights of experts at the step t defined by (4.42)).

9. Using the previous problem and Theorem 4.7 show that the bound $L_T - \min_{i=1,\ldots,N} L_T^i \leqslant \sqrt{T \ln N}$ is valid for the optimal allocation algorithm $Hedge(\beta_t)$ with a variable learning rate $\beta_t = e^{-\eta_t}$, where $\eta_t = \sqrt{\frac{4 \ln N}{t}}$.

10. Prove that for any finite set of N experts which make their forecasts $\xi_t^i \in [0,1]$ at steps $t = 1, 2, \ldots$ and calculate their losses using a loss function: absolute, quadratic or logarithmic, there is a sequence of outcomes for which the cumulative loss of each expert for the first T steps is of the order of $O(T - O(\sqrt{T \ln N})$.

11. Develop a pseudocode of the resampling algorithm which given a sample and a probability distribution generates a new sample whose elements are taken from the given sample and are distributed according to this probability distribution.

Chapter 5

Aggregating algorithm

Machine learning algorithms discussed in Section 4 have regret (training error) $O(\sqrt{T \ln N})$, where T is the length of a learning period, N is the number of experts. In this section we show that for some special loss functions, including square loss and logarithmic loss, this error can be reduced significantly to $O(\ln N)$. In this section the general requirements for such loss functions will be formulated and the corresponding aggregating algorithm with regret $O(\ln N)$ will be presented.

In addition to these properties, aggregating algorithm is an algorithm of a very general nature. In a sense, it performed as well as any other known expert algorithm. Most of the problems that can be solved by such algorithms can also be solved by the aggregating algorithm.

5.1. Mixable loss functions

Consider the simplest case, where the set of outcomes is binary: $\Omega = \{0, 1\}$, and the set of predictions is the unit interval $\Gamma = [0, 1]$. Similarly, we will consider the case of $\Omega = \{-1, 1\}$ and $\Gamma = [-1, 1]$.

We assume that the loss function $\lambda(\omega, \gamma)$ is nonnegative and satisfies the following conditions:

- For any ω, the function $\lambda(\omega, \gamma)$ is continuous by γ;

- A real number $\gamma \in [0,1]$ exists such that both values $\lambda(0,\gamma)$ and $\lambda(1,\gamma)$ are finite;
- There is no $\gamma \in [0,1]$ such that both values $\lambda(0,\gamma)$ and $\lambda(1,\gamma)$ are infinite.

For any loss function $\lambda(\omega,\gamma)$, define *the prediction set*:

$$\Pi_\lambda = \{(x,y) : \exists\ p\ (\lambda(0,p) = x, \lambda(1,p) = y)\} \tag{5.1}$$

and *the superprediction set*

$$\Sigma_\lambda = \{(x,y) : \exists\ p\ (\lambda(0,p) \leqslant x, \lambda(1,p) \leqslant y)\}. \tag{5.2}$$

We call the corresponding half-plane $[0,+\infty)^2$ containing the prediction and superprediction sets *the prediction space*.

The first property of loss function and compactness property of the interval $[0,1]$ imply that the superprediction set is compact.

For functions considered below, the prediction set (5.1) is a boundary of the set (5.2) of superpredictions.

For any $\eta > 0$, let $E_\eta : [0,+\infty)^2 \to (0,1]^2$ be a homomorphism from the prediction space to *the exponential space*

$$E_\eta(x,y) = (e^{-\eta x}, e^{-\eta y}) \tag{5.3}$$

for all $x, y \in [0,+\infty)$.

This homomorphism transforms the prediction set (5.1) to a set

$$E_\eta(\Pi_\lambda) = \{(e^{-\eta\lambda(0,p)}, e^{-\eta\lambda(1,p)}) : p \in \Gamma\},$$

and the superprediction set (5.2) to a set

$$E_\eta(\Sigma_\lambda) = \{(x,y) : \exists\ p\ (0 \leqslant x \leqslant e^{-\eta\lambda(0,p)}, 0 \leqslant y \leqslant e^{-\eta\lambda(1,p)})\}. \tag{5.4}$$

A loss function $\lambda(\omega,\gamma)$ is called *η-mixable*, if the set $E_\eta(\Sigma_\lambda)$ is convex. A loss function is called *mixable*, if it is η-mixable for some $\eta > 0$.

Clearly, for any mixable loss function, the superprediction set is convex. We will see that not every loss function with a convex

superprediction set is mixable. Thus, mixability is a more stronger requirement than just the convexity of the superprediction set.

We will consider logarithmic, square, absolute and simple loss functions. The first two of them are mixable.

Let Ω be a finite set, Γ be a set of all probability distributions on the set Ω. The logarithmic loss function is defined

$$\lambda(\omega, \gamma) = -\ln \gamma\{\omega\},$$

where $\omega \in \Omega$, $\gamma \in \Gamma$ is a probability distribution, and $\gamma(\omega)$ is the probability of an element $\omega \in \Omega$.

In the case $\Omega = \{0, 1\}$ we identify γ with the probability of 1, then $1 - \gamma$ is the probability of 0. In this case $\Gamma = [0, 1]$ and the logarithmic loss function is presented in the form

$$\lambda(\omega, \gamma) = -\ln(\omega\gamma + (1 - \omega)(1 - \gamma)),$$

or, in more detail,

$$\lambda(\omega, \gamma) = \begin{cases} -\ln \gamma \text{ if } \omega = 1, \\ -\ln(1 - \gamma) \text{ if } \omega = 0. \end{cases}$$

A generalized loss function is defined as

$$\lambda(\omega, \gamma) = -\frac{1}{\eta} \ln(\omega\gamma + (1 - \omega)(1 - \gamma)), \tag{5.5}$$

where $\eta > 0$ is a parameter.

The square loss function is defined as

$$\lambda(\omega, \gamma) = c(\omega - \gamma)^2,$$

where c is a positive constant.

We will consider $\Omega = \{0, 1\}$ and $\Gamma = [0, 1]$ or $\Omega = [-1, 1]$ and $\Gamma = [-1, 1]$.

The absolute loss function is

$$\lambda(\omega, \gamma) = c|\omega - \gamma|,$$

where c is a positive constant. The outcomes and prediction sets are the same as for the square loss function.

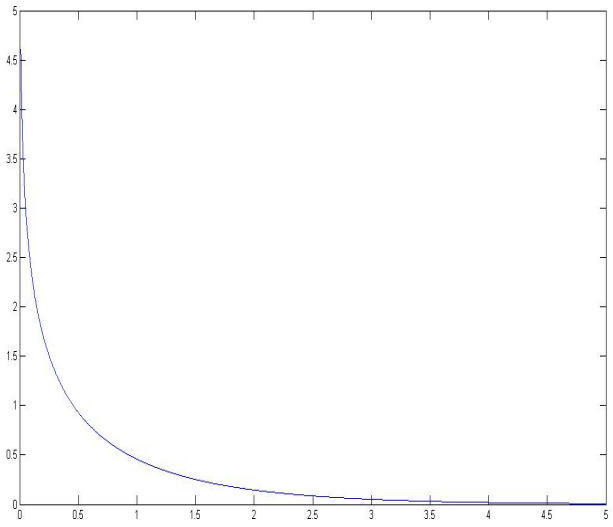


Figure 6.1. The prediction and superprediction sets for the logarithmic loss function

A simple prediction game and the simple loss function are considered in the case where $\Omega = \Gamma = \{0, 1\}$. The simple loss function coincides with the absolute loss function where $c = 1$:

$$\lambda(\omega, \gamma) = \begin{cases} 0 \text{ if } \omega = \gamma, \\ 1 \text{ otherwise.} \end{cases}$$

Let us discuss the geometric properties of the mixable loss functions. Here generalized logarithmic loss function plays a special role.

It is easy to see that the prediction set (5.1) of the generalized logarithmic loss function (5.5) is a curve:

$$\{(x, y) : e^{-\eta x} + e^{-\eta y} = 1\}. \tag{5.6}$$

We will consider parallel shifts of the curve (5.6) in the prediction

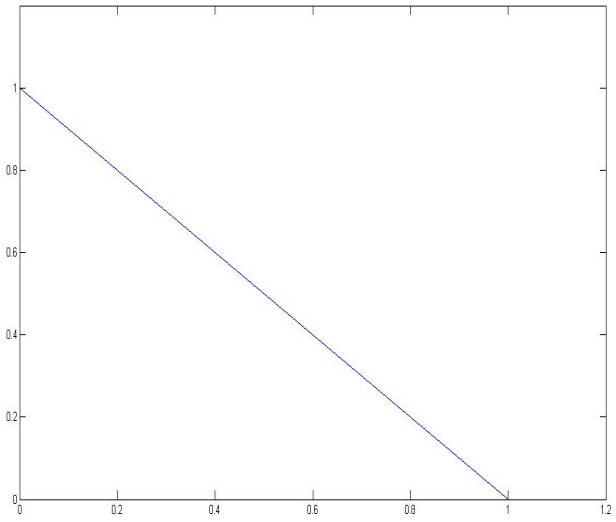


Figure 6.2. The images of the prediction and superprediction set in the exponential space for the logarithmic loss function

half-plane, i.e, all curves of the form

$$\{(x, y) : e^{-\eta(x-\alpha)} + e^{-\eta(y-\beta)} = 1\} \tag{5.7}$$

for any vector (α, β).

We say that a point (x_1, y_1) is located *Northeast* of a point (x_2, y_2) if $x_1 \geqslant x_2$ and $y_1 \geqslant y_2$.

A set $A \subseteq \mathcal{R}^2$ is located *Northeast* of some parallel shift of the curve (5.6) if every its point is located *Northeast* of some point located on the shift (5.7).

Note that all parallel shifts of the curve $e^{-\eta x} + e^{-\eta y} = 1$ in the exponential space coincide with preimages of all straight lines $ax + by = c$ considered in the prediction space, where $a > 0$ and $b > 0$. Indeed, it is easy to verify that a preimage of the straight line

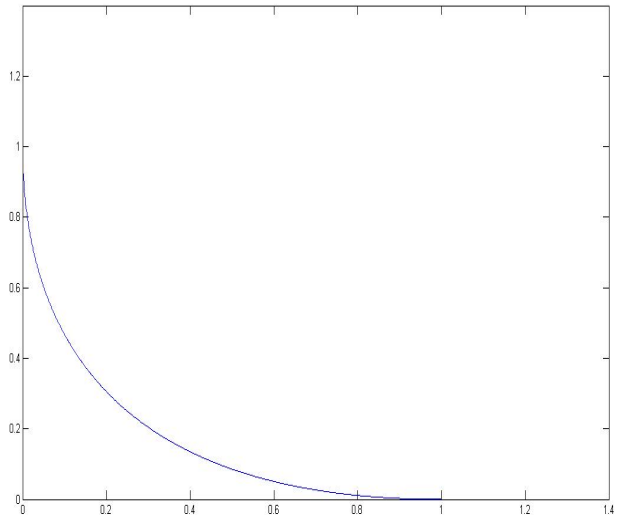


Figure 6.3. The prediction and superprediction sets for the square loss function

$ax + by = c$ under homomorphism E_η is a curve

$$ae^{-\eta x} + be^{-\eta y} = c,$$

that is a parallel shift of the curve $e^{-\eta x} + e^{-\eta y} = 1$ defined by the vector

$$\left(-\frac{1}{\eta}\ln\frac{a}{c}, -\frac{1}{\eta}\ln\frac{b}{c}\right).$$

Thus, there is a one-to-one correspondence between all such straight lines $ax + by = c$ considered in the exponential space and all parallel shifts of the curve $e^{-\eta x} + e^{-\eta y} = 1$ considered in the prediction space.

It is easy to see that the image $E_\eta(\Sigma_\lambda)$ of the superprediction set in the exponential space is convex if and only if for every point of its boundary there is a straight line passing through this point so that

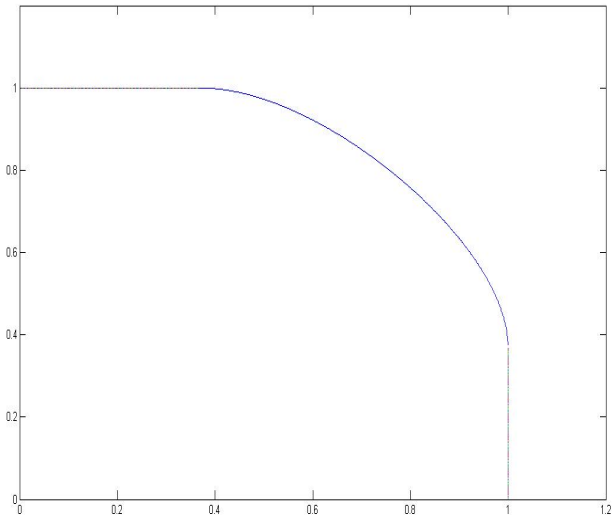


Figure 6.4. The images of the prediction and superprediction sets in the exponential space for the square loss function

the whole image of superprediction set is located at one side of this line.

Transferring this property from the the exponential space to he prediction space, we obtain the following characteristic property of mixability of the loss function.

Proposition 5.1. *A loss function is η-mixable if and only if for any point (a, b) locating at the boundary of the superprediction set a parallel shift $e^{-\eta(x-\alpha)} + e^{-\eta(y-\beta)} = 1$ of the curve $e^{-\eta x} + e^{-\eta y} = 1$ exists passing through the point (a, b) and such that the whole superprediction set lies Northeast of this shift.*

In the following sections we will consider the mixable loss functions. It turns out that at certain intervals of values η logarithmic and square loss functions are η-mixable; the absolute loss function

has not this property. Highly effective in this case is the so-called aggregating algorithm, which was discovered by Vovk [35] in 1990. Hystorically, it is one of the first of averaging algorithms of this kind. This algorithm is a generalization of a more simple weighted majority algorithm which was proposed in 1989 by Littlestone and Varmuth [22].

Aggregating algorithm has prediction error, which depends only on the number of experts and does not depend on the length of the outcome sequence.

5.2. Finite set of experts

The prediction algorithms constructing in Sections 4.4 and 4.6 have regret of order $O(\sqrt{T \ln N})$, where T is the length of prediction period and N is the number of experts. Algorithms and results of that sections refer to loss function of arbitrary form with an exception that in Section 4.4 we require convexity of a loss function by a forecast.

In this section, we present a weighting algorithm, which has an optimal regret for any loss function, and for a mixable loss function this algorithm has regret $O(\ln N)$ independent of the length T of the prediction period, where N is the number of experts.

In general, the bound of the loss of aggregating algorithm has the form

$$L_T \leqslant c(\eta) \inf_{\theta} L_T(\theta) + a(\eta) \ln N$$

for all T. For a mixable loss function, $c(\eta) = 1$ for some values of the learning rate η.

We first consider a scheme of the algorithm in the case where the set of outcomes is $\Omega = \{0, 1\}$ and a set of experts $\Theta = \{1, 2, \ \ldots, \ N\}$ is finite. The prediction set is Γ. Let a loss function $\lambda(\omega, \gamma)$ be given, where $\omega \in \Omega$ and $\gamma \in \Gamma$.

We consider in the next sections infinite (and even uncountable) sets of expert Θ. In this case, the results did not significantly change. We only introduce measures on experts and replace sum of experts on the integrals over θ and $\gamma \in \Gamma$.

Recall the protocol of a game of prediction with expert advice.

Let $L_0 = 0$, $L_0(i) = 0$, $i = 1, \ \ldots, \ N$.
FOR $t = 1, 2, \ \ldots$
Expert i announces a forecast $\xi_t^i \in \Gamma$, $i = 1, \ \ldots, \ N$.
Learner announces a forecast $\gamma_t \in \Gamma$.
Nature announces an outcome $\omega_t \in \Omega$.
Expert i updates its cumulative loss at step t:

$$L_t(i) = L_{t-1}(i) + \lambda(\omega_t, \xi_t^i).$$

Learner updates its cumulative loss at step t:

$$L_t = L_{t-1} + \lambda(\omega_t, \gamma_t).$$

ENDFOR

Fix the learner rate $\eta > 0$ and define $\beta = e^{-\eta}$. We introduce a priori distribution $P_0(i)$ on the set of experts Θ. It is natural to take the uniform a priori probability distribution on set of experts $P_0(i) = 1/N$ for all $i \in \Theta$, where N is the number of experts.

Learner updates the experts weights on steps $t = 1, 2, \ \ldots$ by a rule:

$$P_t(i) = \beta^{\lambda(\omega_t, \xi_t^i)} P_{t-1}(i), \tag{5.8}$$

where $i = 1, \ \ldots, \ N$. Therefore, the weight of an expert suffering large loss is reduced.

The experts weights are normalized:

$$P_t^*(i) = \frac{P_t(i)}{\sum\limits_{j=1}^{N} P_t(j)} \tag{5.9}$$

such that their sum becomes equal to 1.

Consider an auxiliary function which is called "pseudoprediction":

$$g_t(\omega) = \log_\beta \sum_{i=1}^{N} \beta^{\lambda(\omega, \xi_t^i)} P_{t-1}^*(i). \tag{5.10}$$

We call the formulae (5.10) *Aggregating Pseudo Algorithm* and denote it APA. Define the cumulative loss of the APA algorithm at first T

steps on a sequence $\omega_1, \ \ldots, \ \omega_T$ of outcomes

$$L_T(APA) = \sum_{t=1}^{T} g_t(\omega_t). \tag{5.11}$$

The following lemma represents the cumulative loss of the APA algorithm in a more convenient way.

Lemma 5.1. *The cumulative loss of the APA algorithm at first T steps is equal to*

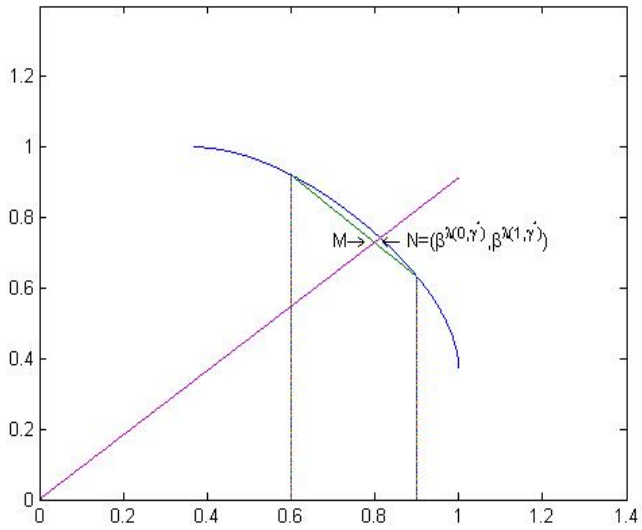
$$L_T(APA) = \log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i).$$

Proof. By (5.8) we obtain

$$P_T(i) = \beta^{\sum\limits_{t=1}^{T} \lambda(\omega_t, \xi_t^i)} P_0(i) = \beta^{L_T(i)} P_0(i).$$

By definition we have

$$\begin{aligned}
\log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i) - \log_\beta \sum_{i=1}^{N} \beta^{L_{T-1}(i)} P_0(i) = \\
= \log_\beta \frac{\sum\limits_{i=1}^{N} \beta^{L_T(i)} P_0(i)}{\sum\limits_{i=1}^{N} \beta^{L_{T-1}(i)} P_0(i)} = \\
= \log_\beta \frac{\sum\limits_{i=1}^{N} \beta^{L_{T-1}(i) + \lambda(\omega_T, \xi_T^i)} P_0(i)}{\sum\limits_{i=1}^{N} \beta^{L_{T-1}(i)} P_0(i)} = \\
= \log_\beta \frac{\sum\limits_{i=1}^{N} \beta^{\lambda(\omega_T, \xi_T^i)} P_{T-1}(i)}{\sum\limits_{i=1}^{N} P_{T-1}(i)} = \\
= \log_\beta \sum_{j=1}^{N} \beta^{\lambda(\omega_T, \xi_T^j)} P_{T-1}^*(j) = g_T(\omega_T).
\end{aligned} \tag{5.12}$$



. 6.5. The calculation of the prediction γ^*. The straight line passing through the point M marks a point on the curve which is used to calculate the prediction γ^*.

The last equality followws from (5.10). Since the equality (5.12) holds for all T, we obtain the assertion of the lemma: $L_T(APA) = \sum_{t=1}^{T} g_t(\omega_t) = \log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i)$. $\triangle$

The pseudoprediction $g_t(\omega)$ represents some average loss and does not give the prediction $\gamma \in \Gamma$ itself for which this loss is evaluated.

In some cases, a superprediction can be transformed into a prediction. *A substitution function* is a function $\gamma_t = \Sigma(g_t)$ such that $\lambda(\omega, \Sigma(g_t)) \leqslant g_t(\omega)$ for all ω.

We will show that a substitution function exists if a loss function $\lambda(\omega, \gamma_t)$ is mixable.

Proposition 5.2. *If a loss function is mixable then a substitution*

function exists.

Proof. Assume that a loss function $\lambda(\omega, \gamma)$ is η-mixable and $\beta = e^{-\eta}$. Since the image

$$E_\eta(\Sigma_\lambda) = \{(x, y) : \exists\ p\ (0 \leqslant x \leqslant \beta^{\lambda(0,p)} 0 \leqslant y \leqslant \beta^{\lambda(1,p)})\}$$

of the superprediction set of the loss function $\lambda(\omega, \gamma)$ is convex, an $\gamma^* \in \Gamma$ exists such that

$$\beta^{\lambda(\omega_T, \gamma^*)} \geqslant \sum_{j=1}^{N} \beta^{\lambda(\omega_T, \xi_T^j)} P_{T-1}^*(j) \tag{5.13}$$

for all $\omega_T \in \{0, 1\}$.

The inequality (5.13) means that the abscissa and the ordinate of the point

$$\left(\beta^{\lambda(0,\gamma^*)}, \beta^{\lambda(1,\gamma^*)}\right)$$

are more than or equal to the abscissa and the ordinate of the point

$$\left(\sum_{j=1}^{N} \beta^{\lambda(0,\xi_T^j)} P_{T-1}^*(j), \sum_{j=1}^{N} \beta^{\lambda(1,\xi_T^j)} P_{T-1}^*(j)\right),$$

correspondently. Define $\Sigma(g_t) = \gamma^*$. The condition $\lambda(\omega, \Sigma(g_t)) \leqslant g_t(\omega)$ holds for all ω.

Given an algorithm computing loss function $\lambda(\omega, \gamma)$, an algorithm computing the prediction $\gamma_t = \Sigma(g_t)$ can be easily constructed. Such an algorithm is called *Aggregating Algorithm* or AA algorithm.

In case, where a function $\Sigma(g_t)$ exists, by Lemma 5.1, the inequality

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma(g_t)) \leqslant$$

$$\leqslant L_T(APA) = \log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i) \tag{5.14}$$

holds.

Assign the equal weights to all experts: $P_0(i) = 1/N$. Then by (5.14) we have for any $i \in \Theta$

$$L_T(AA) \leqslant \log_\beta \left(\frac{1}{N} \sum_{i=1}^{N} \beta^{L_T(i)} \right) \leqslant$$
$$\leqslant \log_\beta \left(\frac{1}{N} \beta^{L_T(i)} \right) = L_T(i) + \frac{\ln N}{\eta} \tag{5.15}$$

for all T.

The bound (5.15) means that the cumulative loss of the aggregating algorithm AA does not exceed the total loss of any expert, including the best expert that has the lowest total loss among all experts up to some regret. It is important that this regret depends only on the number of experts and does not depend on the length of the prediction period as it did for the exponential weighting algorithms.

5.3. Infinite set of experts

We reproduce the scheme of the algorithm AA for the case of an infinite number of experts Θ. We assume that the set Θ is endowed with the structure of probability space – a sigma algebra of Borel sets and a probability measure on it. In this case the sums of experts are replaced on integrals by these measures.

As usual, $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$. A loss function $\lambda(\omega, \gamma)$ be given, where $\omega \in \Omega$ and $\gamma \in \Gamma$. Let $\eta > 0$ be a learning rate and $\beta = e^{-\eta}$. Let also, some a priory probability distribution $P_0(d\theta)$ on the set of all experts Θ be given.

Learner updates the experts weights on step t by the rule

$$P_t(d\theta) = \beta^{\lambda(\omega_t, \xi_t^\theta)} P_{t-1}(d\theta). \tag{5.16}$$

Therefore, the weight of any expert which suffer greater cumulative loss decreases.

By definition the rule (5.16) is equivalent to the rule

$$P_t(E) = \int_E \beta^{\lambda(\omega_t, \xi_t^\theta)} P_{t-1}(d\theta),$$

where E is an arbitrary event.

We normilize the weights (5.16) of experts:

$$P_t^*(d\theta) = \frac{P_t(d\theta)}{P_t(\Theta)}. \tag{5.17}$$

These normalized weights define a probability distribution for that $P_t^*(\Theta) = 1$.

Similarly, define the pseudoprediction

$$g_t(\omega) = \log_\beta \int_\Theta \beta^{\lambda(\omega,\xi_t^\theta)} P_{t-1}^*(d\theta). \tag{5.18}$$

The algorithm computing the pseudoprediction is also denoted APA. Its cumulative loss for T steps is equal to

$$L_T(APA) = \sum_{t=1}^{T} g_t(\omega_t). \tag{5.19}$$

By (5.16),

$$P_T(d\theta) = \beta^{\sum_{t=1}^{T} \lambda(\omega_t,\xi_t^\theta)} P_0(d\theta) = \beta^{L_T(\theta)} P_0(d\theta),$$
$$P_T^*(d\theta) = \frac{\beta^{L_T(\theta)}}{\int_\Theta \beta^{L_T(\theta)} P_0(d\theta)} P_0(d\theta).$$

We rewrite the equality (5.18) in the form

$$g_T(\omega) = \log_\beta \int_\Theta \frac{\beta^{\lambda(\omega,\xi_T^\theta)+L_{T-1}(\theta)}}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} P_0(d\theta). \tag{5.20}$$

An analogue of Lemma 5.1 holds:

Lemma 5.2. *The cumulative loss of the APA algorithm for T steps can be represented in the form*

$$L_T(APA) = \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta). \tag{5.21}$$

Proof. The proof of this lemma is similar to the proof of Lemma 5.1. By (5.8), we obtain

$$P_t(d\theta) = \beta^{\sum_{t=1}^{T} \lambda(\omega_t, \xi_t^\theta)} P_0(d\theta) = \beta^{L_T(\theta)} P_0(d\theta). \tag{5.22}$$

Also, by definition we have

$$\begin{aligned}
\log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta) - \log_\beta \int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta) = \\
= \log_\beta \frac{\int_\Theta \beta^{L_T(\theta)} P_0(d\theta)}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} = \\
= \log_\beta \frac{\int_\Theta \beta^{L_{T-1}(\theta) + \lambda(\omega_T, \xi_T^\theta)} P_0(d\theta)}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} = \\
= \log_\beta \frac{\int_\Theta \beta^{\lambda(\omega_T, \xi_T^\theta)} P_{T-1}(d\theta)}{\int_\Theta P_{T-1}(d\theta)} = \\
= \log_\beta \int_\Theta \beta^{\lambda(\omega_T, \xi_T^\theta)} P_{T-1}^*(d\theta) = g_T(\omega_T).
\end{aligned} \tag{5.23}$$

The last inequality follows from the definition (5.18).

Since (5.23) holds for all T, we obtain the assertion of the lemma. $\triangle$

It is easy to show that in the case of infinite experts space Θ and mixable loss function the substitution function $\Sigma(g_t)$ also exists. Indeed, the integrals over $d\theta$ can be approximated by finite sums, which correspond to finite sets of experts. Since the set of predictions is compact, the predictions computed using the AA algorithm for these finite sets of experts, have a limit point γ^*. Since the loss function $\lambda(\omega, \gamma)$ is continuous by γ, this limit point satisfies

$$\lambda(\omega, \gamma^*) \leqslant g_t(\omega)$$

for all ω, where $g_t(\omega)$ is defined by (5.18). Put $\Sigma(g_t) = \gamma^*$.

Then by Lemma 5.2,

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma(g_t)) \leqslant \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta). \tag{5.24}$$

5.4. Arbitrary loss function

We will show in the following sections that the logarithmic and square loss functions are mixable.

In general case, for a non-mixable loss function, *a mixability curve* $c(\eta)$ is defined:

$$c(\eta) = \inf \left\{ c : \forall\ P\ \exists\ \delta \in \Gamma\ \forall\ \omega \left(\lambda(\omega, \delta) \leqslant c \log_\beta \int_\Gamma \beta^{\lambda(\omega,\gamma)} P(d\gamma) \right) \right\}.$$

Under some natural assumptions on the initial sets the function $c(\eta)$ is continuous and non-increasing.

The substitution function is defined as a function satisfying the condition

$$\forall\ \omega : \lambda(\omega, \Sigma_\eta(g)) \leqslant c(\eta) g(\omega) \tag{5.25}$$

for each pseudoprediction function

$$g(\omega) = \log_\beta \int_\Gamma \beta^{\lambda(\omega,\gamma)} P(d\gamma)$$

and probability distribution P on Γ.

A minimax substitution function can be defined:

$$\Sigma_\eta(g) \in \arg\min_{\gamma\in\Gamma} \sup_{\omega\in\Omega} \frac{\lambda(\omega,\gamma)}{g(\omega)}. \tag{5.26}$$

By definition any minimax substitution function $\Sigma_\eta(g)$ defined by (5.26) also satisfies (5.25).

Note that there could be other - not minimax substitution functions such that the condition (5.25) holds. Often they are easier to calculate.

In general case, for any loss function not necessary mixable, we have

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma_\eta(g_t)) \leqslant c(\eta) \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta) \tag{5.27}$$

instead of (5.24). The similar inequalities hold if we introduce the factor $c(\eta)$.

In the case of a finite experts set the inequality (5.15) becomes

$$L_T(AA) \leqslant c(\eta)\log_\beta\left(\frac{1}{N}\sum_{i=1}^{N}\beta^{L_T(i)}\right) \leqslant$$

$$\leqslant c(\eta)\log_\beta\left(\frac{1}{N}\beta^{L_T(i)}\right) = c(\eta)L_T(k) + c(\eta)\frac{\ln N}{\eta}$$

for all T and all $k = 1,\ \ldots,\ N$.

5.5. Logarithmic loss function

Assume that a set Ω of outcomes and a set Θ of experts are finite and a prediction set $\Gamma = \mathcal{P}(\Omega)$ be the set of all probability distributions on Ω. For any $\gamma \in \Gamma$, let $\gamma(\omega) = \gamma(\{\omega\})$ be the probability of $\omega \in \Omega$. *The logarithmic loss* function is defined $\lambda(\omega, \gamma) = -\ln\gamma(\omega)$.

Put $\eta = 1$ and $\beta = e^{-1}$. In this case

$$\beta^{\lambda(\omega,\gamma)} = \gamma(\omega)$$

is equal to the probability that an expert or *Learner* assigns to an outcome ω. In this case, aggregating algorithm coincides with the algorithm the exponential weighting.

An *Expert* i prediction at step t is a probability distribution $\xi_t^i = \xi_t^i(\cdot) \in \Gamma$ on the set of outcomes Ω.

At any step t, we assign with an expert $i \in \Theta$ a probability distribution Q_i on the set Ω^∞ defined by the conditional probabilities:

$$Q_i(\omega|\omega_1,\ \ldots,\ \omega_{t-1}) = \xi_t^i(\omega) \in \Gamma. \tag{5.28}$$

We can interpreted this probability distribution as a subjective conditional distribution of *Expert* i at step t. More correctly, the value (5.28) is equal to the conditional probability which *Expert* i assigns to an a future outcome ω after observing past outcomes $\omega_1, \ldots,\ \omega_{t-1}$. Accordingly, a subjective probability assigned at a step t by *Expert* i to the whole sequence of outcomes $\omega_1,\ \ldots,\ \omega_t$ is equal to the product of these conditional probabilities

$$Q_i(\omega_1,\ \ldots,\ \omega_t) = \xi_1^i(\omega_1)\xi_2^i(\omega_2)\cdot\ \ldots\ \cdot\xi_t^i(\omega_t). \tag{5.29}$$

The experts weights are updated according to the rule (5.8). For logarithmic loss function, the weight of *Expert* i is updated at step t:

$$\begin{aligned} P_t(i) = \beta^{\lambda(\omega_t,\xi_t^i)} P_{t-1}(i) = \\ = \xi_1^i(\omega_1)\xi_2^i(\omega_2) \cdot \ \ldots \ \cdot \xi_t^i(\omega_t) P_0(i) = \\ = Q_i(\omega_1, \ \ldots, \ \omega_t) P_0(i). \end{aligned} \tag{5.30}$$

The weights (5.30) of experts are normalized by the formulae

$$P_t^*(i) = \frac{P_t(i)}{\sum\limits_{j=1}^{N} P_t(j)} = \frac{Q_i(\omega_1, \ \ldots, \ \omega_t) P_0(i)}{\sum\limits_{j=1}^{N} Q_j(\omega_1, \ \ldots, \ \omega_t) P_0(j)}. \tag{5.31}$$

The probability $P_t^*(i)$ is called *a posterior probability* of the expert i after observing the outcomes $\omega_1, \ \ldots, \ \omega_t$.

Since $\beta^{\lambda(\omega_t,\xi_t^i)} = \xi_t^i(\omega_t)$, the pseudoprediction (5.10) is equal to the logarithm of the Bayesian mixture of the probability distributions presented by experts at step t:

$$g_t(\omega) = \log_\beta \sum_{i=1}^{N} \xi_t^i(\omega) P_{t-1}^*(i). \tag{5.32}$$

In this case, the prediction $\Sigma(g_t)$ of the aggregating algorithm is equal to the Bayesian mixture of probability distributions presented by experts at step t. This probability distribution is defined by the rule

$$\gamma_t(\omega) = \Sigma(g_t) = \sum_{i=1}^{N} \xi_t^i(\omega) P_{t-1}^*(i).$$

The value of the logarithmic loss function on an outcome ω_t and on a *Learner's* forecast γ_t is equal to the pseudoprediction

$$\lambda(\omega_t, \gamma_t) = -\ln \gamma_t(\omega_t) = \log_\beta \sum_{i=1}^{N} \xi_t^i(\omega_t) P_{t-1}^*(i) = g_t(\omega_t).$$

We explain this method and its relationship with the *Bayesian rule* in more details on example of the first two steps: $t = 1, 2$.

At step 1, each expert i outputs its forecast $\xi_1^i = \xi_1^i(\cdot) \in \Gamma$ that is a probability distribution on Ω. Then the forecast of the algorithm AA is computed as a Bayesian mixture of probability distributions of the experts with respect to the a priory distribution P_0 on the set of experts:

$$\gamma_1(\omega) = \sum_{i=1}^{N} \xi_1^i(\omega) P_0(i).$$

Once a first outcome ω_1 is revealed by *Nature*, *Learner* updates the prior distribution on the set of experts as follows. At first, *Learner* defines the expert weights:

$$P_1(i) = \beta^{\lambda(\omega_1, \xi_1^i)} P_0(i) = \xi_1^i(\omega_1) P_0(i).$$

After that, *Learner* normalizes these weights and obtains the posterior probabilities of experts after observing the outcome ω_1:

$$P_1^*(i) = \frac{\xi_1^i(\omega_1) P_0(i)}{\sum\limits_{j=1}^{N} \xi_1^j(\omega_1) P_0(j)}.$$

It is easy to see that this formula is the well known *Bayesian rule* for computing the posterior probability $P_1^*(i)$ of any expert i after observing the outcome ω_1.

The same is true at step $t = 2$. At step $t = 2$, each expert i outputs a forecast – a probability distribution $\xi_2^i(\cdot)$ on Ω. The forecast of the aggregating algorithm AA is computed by the rule

$$\gamma_2(\omega) = \sum_{i=1}^{N} \xi_2^i(\omega) P_1^*(i)$$

that is a Bayesian mixture of the experts probability distributions with respect to the posterior probability P_1^* on the set of experts computed at the step $t = 1$.

After receiving the second outcome ω_2, *Learner* updates the posterior probability on the set of experts as follows. At first, he updates the experts weights

$$P_2(i) = \beta^{\lambda(\omega_2, \xi_2^i)} P_1(i) = \xi_2^i(\omega_2) P_1(i) = \xi_1^i(\omega_1) \xi_2^i(\omega_2) P_0(i)$$

and, after that, he computes the new posterior probabilities of the experts normalizing their weights:

$$P_2^*(i) = \frac{\xi_1^i(\omega_2)P_1^*(i)}{\sum\limits_{j=1}^{N} \xi_2^j(\omega_2)P_1^*(j)}.$$

We can see again that this formula is the *Bayesian rule* for computing the posterior probability $P_2^*(i)$ of any expert i after observing the outcomes ω_1, ω_2.

Thus, in the case of the logarithmic loss function the aggregating algorithm AA can be represented as the online Bayesian method.

The loss of the ith expert over T steps is equal to

$$\begin{aligned} L_T(i) &= \sum_{i=1}^{T} \lambda(\omega_t, \xi_t^i) = \\ &= -\ln(\xi_1^i(\omega_1) \cdot \ \ldots \ \cdot \xi_T^i(\omega_T)) = \\ &= -\ln Q_i(\omega_1, \ldots, \ \omega_T). \end{aligned} \tag{5.33}$$

This equality is a subjective probability (5.28) assigned by *Statistician* to the expert i at step t.

The cumulative loss of *Learner* over the first T steps is equal to

$$\begin{aligned} L_T(AA) &= \sum_{t=1}^{T} \lambda(\omega_t, \Sigma(g_t)) = \\ &= \log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i) = \\ &= \log_\beta \sum_{i=1}^{N} Q_i(\omega_1, \ldots, \ \omega_T) P_0(i). \end{aligned} \tag{5.34}$$

Thus, the cumulative loss of *Learner* at first T steps is equal to the minus logarithm of the Bayesian mixture of probabilities assigned by the experts to the sequence of outcomes $\omega_1, \ldots, \ \omega_T$ of length T.

The inequality (5.15) is transformed to the inequality

$$L_T(AA) = \log_\beta \sum_{i=1}^{N} Q_i(\omega_1, \dots, \omega_T) P_0(i) \leqslant$$
$$\leqslant -\ln Q_k(\omega_1, \dots, \omega_T) - \ln P_0(k) \tag{5.35}$$

for all T and $k = 1, \dots, N$.

5.6. Simple prediction game

Recall that the simple prediction game is valid for binary outcome and prediction spaces: $\Omega = \Gamma = \{0, 1\}$. The prediction task for this game is precisely to predict the future outcome. The loss function is defined

$$\lambda(\omega, \gamma) = \begin{cases} 0 \text{ if } \omega = \gamma, \\ 1 \text{ otherwise.} \end{cases}$$

Therefore, the cumulative loss of an expert or the learner is equal to the total numbers of prediction errors.

Assume that there are N experts. An expert i outputs a forecast $\xi_t^i \in \{0, 1\}$ at any step t.

We present any pseudoprediction

$$g(\omega) = \log_\beta \sum_{i=1}^{N} \beta^{\lambda(\omega, \xi_t^i)} P_{t-1}^*(i) \tag{5.36}$$

as a point $(g(0), g(1))$ at the positive half-plane. This point can be written in the form

$$(\log_\beta(\beta p + (1-p)), \ \log_\beta(p + \beta(1-p))), \tag{5.37}$$

where $0 < \beta < 1$ is a learning rate, $p = \sum\limits_{\xi_t^i = 1} P_{t-1}^*(i)$ is the total weight of all experts i predicting $\xi_t^i = 1 \quad t$, and $1 - p = \sum\limits_{\xi_t^i = 0} P_{t-1}^*(i)$ is the total weight of all experts i predicting $\xi_t^i = 0 \quad t$

The points (5.37) form a convex curve connecting the points $(1, 0)$ and $(0, 1)$ corresponding to $p = 0$ and $p = 1$.

By definition $1/c(\beta)$ is equal to the abscissa (ordinate) of the point that is an intersection of the line $y = x$ and the curve.

By (5.37) we obtain for $p = \frac{1}{2}$

$$\frac{1}{c(\beta)} = \log_\beta \left(\frac{1+\beta}{2} \right),$$

or

$$c(\beta) = \frac{\ln \frac{1}{\beta}}{\ln \frac{2}{1+\beta}}. \tag{5.38}$$

We apply the aggregating algorithm AA for this game. Define the substitution function $\gamma = \Sigma(g)$ as follows: $\Sigma(g) = 0$ if the point $(g(0), g(1))$ computed by (5.37) lies higher than the stright line $y = x$, define $\gamma = \Sigma(g) = 1$ if the point $(g(0), g(1))$ lies below or on the stright line $y = x$.

This substitution function satisfies the condition (5.25), since for $\gamma = 0$ it holds for the abscissa $g(0) \geqslant \lambda(0, 0) = 0$ and for the ordinate $g(1) \geqslant \frac{1}{c(\beta)}$ of the intersection point of the bisector of the coordinate angle and the curve (5.37). Then $g(1) \geqslant \frac{1}{c(\beta)} = \frac{1}{c(\beta)}\lambda(1, 0)$. Therefore, $\lambda(\omega, 0) \leqslant c(\beta)g(\omega)$ for all $\omega \in \{0, 1\}$.

Similarly, we obtain the inequality $\lambda(\omega, 1) \leqslant c(\beta)g(\omega)$ for all $\omega \in \{0, 1\}$.

Note that if the point $(g(0), g(1))$ lies higher than the line $y = x$ then the abscissa is less than ordinates, ie, $g(0) < g(1)$ or

$$\log_\beta(\beta p + (1-p)) < \log_\beta(p + \beta(1-p))$$

that is equivalent to $p < \frac{1}{2}$. In this case the algorithm predicts $\gamma = 0$.

Otherwise, if the point $(g(0), g(1))$ lies below or on the stright line $y = x$ then

$$\log_\beta(\beta p + (1-p)) \geqslant \log_\beta(p + \beta(1-p)),$$

that is equivalent to $p \geqslant \frac{1}{2}$. In this case the algorithm predicts $\gamma = 1$.

This means that the aggregating algorithm predicts $\gamma = 1$ if the total weight of all experts predicting 1 is more than the total weight

of all experts predicting 0; the algorithm AA predicts $\gamma = 0$ otherwise. Thus, the aggregating algorithm AA predicts as the weighted majority algorithm defined in Section (4.1).

In this case for any expert $\theta \in \Theta$ the following inequality holds:

$$L_T(AA) \leqslant \left(\frac{\ln \frac{1}{\beta}}{\ln \frac{2}{1+\beta}} \right) L_T(\theta) + \ln \left(\frac{1}{\ln \frac{2}{1+\beta}} \right) \ln N \tag{5.39}$$

5.7. Square loss function

In this section we study the application of the aggregating algorithm for square loss function in the simplest case, where the outcome set is binary $\Omega = \{-1, 1\}$ and the prediction set is the interval $\Gamma = [-1, 1]$. The square loss function is $\lambda(\omega, \gamma) = (\omega - \gamma)^2$. All results presented below also hold for $\Omega = [-1, 1]$.

Lemma 5.3. *In case $\Omega = \{-1, 1\}$ and $\Gamma = [-1, 1]$ the square loss function is η-mixable if and only if $\eta \leqslant \frac{1}{2}$.*

Proof. We present the pseudoprediction $(g(-1), g(1))$ as a point

$$(e^{-\eta g(-1)}, e^{-\eta g(1)})$$

in the exponential space.

The set of all predictions $\gamma \in [-1, 1]$ defines the paremetric curve in the exponential space

$$(x(\gamma), y(\gamma)) = (e^{-\eta(-1-\gamma)^2}, e^{-\eta(1-\gamma)^2}).$$

A loss function is η-mixable if the image of the superprediction set in the exponential space a convex set, ie, if and only if its bounding curve turns to the left with an increase in γ (in this case the abscissa decreases). This will happens if the following concavity condition of a curve holds: $\frac{d^2y}{d^2x} \leqslant 0$.

Let us compute the second derivative of a parametrically defined curve:

$$\frac{d^2y}{d^2x} = \frac{d\gamma}{dx} \frac{x'(\gamma)y''(\gamma) - x''(\gamma)y'(\gamma)}{(x'(\gamma))^2}. \tag{5.40}$$

With an increase in the parameter γ the value $x(\gamma)$ is decreasing, so $\frac{d\gamma}{dx} < 0$.

The game is η-mixable if and only if $\frac{d^2y}{d^2x} \leqslant 0$ that is equivalent to the condition

$$x'(\gamma)y''(\gamma) - x''(\gamma)y'(\gamma) \geqslant 0.$$

Compute the following derivatives by the parameter γ:

$$\begin{aligned} x'(\gamma) &= -2\eta(1+\gamma)e^{-\eta(1+\gamma)^2}, \\ x''(\gamma) &= 2\eta(-1+2\eta(1+\gamma)^2)e^{-\eta(1+\gamma)^2}, \\ y'(\gamma) &= 2\eta(1-\gamma)e^{-\eta(1-\gamma)^2}, \\ y''(\gamma) &= 2\eta(-1+2\eta(1-\gamma)^2)e^{-\eta(1-\gamma)^2}. \end{aligned} \tag{5.41}$$

The condition of η-mixability requires that for all values of $\gamma \in [-1, 1]$ the following equivalent inequalities hold

$$\begin{aligned} -(1+\gamma)(-1+2\eta(1-\gamma)^2) - \\ -(1-\gamma)(-1+2\eta(1+\gamma)^2) \geqslant 0, \\ \eta(1-\gamma^2) \leqslant \frac{1}{2}, \\ \eta \leqslant \frac{1}{2}. \end{aligned} \tag{5.42}$$

Lemma is proved. $\triangle$

We now find some form of the substitution function $\Sigma(g)$ in the case $\Omega = \{-1, 1\}$ and a finite number of experts $\Theta = \{1, \ \ldots, \ N\}$. Let $\eta = \frac{1}{2}$ and $\beta = e^{-\frac{1}{2}}$. A pseudoprediction

$$g_t(\omega) = \log_\beta \sum_{i=1}^{N} \beta^{\lambda(\omega, \xi_t^i)} P_{t-1}^*(i) \tag{5.43}$$

is defined by the point

$$\begin{aligned} (e^{-\frac{1}{2}g(-1)}, e^{-\frac{1}{2}g(1)}) = \\ = \left(\sum_{i=1}^{N} \beta^{\lambda(-1, \xi_t^i)} P_{t-1}^*(i), \sum_{i=1}^{N} \beta^{\lambda(1, \xi_t^i)} P_{t-1}^*(i) \right), \end{aligned} \tag{5.44}$$

which is located under the concave curve

$$\left(\beta^{\lambda(-1,\gamma)}, \beta^{\lambda(1,\gamma)}\right), \tag{5.45}$$

for $\gamma \in [-1.1]$.

Draw a straight line passing through origin of coordinates and the point (5.44). The slope of this line is equal to

$$k = \frac{\beta^{g(1)}}{\beta^{g_t(-1)}} = e^{\frac{1}{2}g_t(-1)-\frac{1}{2}g_t(1)}. \tag{5.46}$$

The intersection point $\left(\beta^{\lambda(-1,\gamma^*)}, \beta^{\lambda(1,\gamma^*)}\right)$ of this line and the curve (5.45) has the abscissa and ordinate not less than the abscissa and ordinate of the point (5.44) :

$$\begin{aligned} \beta^{\lambda(-1,\gamma^*)} &\geqslant \beta^{g_t(-1)}, \\ \beta^{\lambda(1,\gamma^*)} &\geqslant \beta^{g_t(1)}. \end{aligned} \tag{5.47}$$

An equivalent form of (5.47) is

$$\begin{aligned} \lambda(-1,\gamma^*) &\leqslant g_t(-1), \\ \lambda(1,\gamma^*) &\leqslant g_t(1). \end{aligned} \tag{5.48}$$

Let us compute a forecast γ^*. We find γ^* from the equation

$$\frac{\beta^{g_t(1)}}{\beta^{g_t(-1)}} = \beta^{g_t(1)-g_t(-1)} = \frac{\beta^{\lambda(1,\gamma^*)}}{\beta^{\lambda(-1,\gamma^*)}} = \beta^{\lambda(1,\gamma^*)-\lambda(-1,\gamma^*)}. \tag{5.49}$$

It remains to find the root of this equation

$$\lambda(1,\gamma^*) - \lambda(-1,\gamma^*) = (1-\gamma^*)^2 - (-1-\gamma^*)^2 = g_t(1) - g_t(-1),$$

which is equal to

$$\gamma^* = \frac{1}{4}(g_t(-1) - g_t(1)). \tag{5.50}$$

In more detail, at any step t, we compute the forecast

$$\gamma_t^* = \frac{1}{4}\left(\log_\beta \sum_{i=1}^{N} \beta^{\lambda(-1,\xi_t^i)} P_{t-1}^*(i) - \log_\beta \sum_{i=1}^{N} \beta^{\lambda(1,\xi_t^i)} P_{t-1}^*(i)\right)$$

or

$$\gamma_t^* = -\frac{1}{2}\ln\left(\frac{\sum\limits_{i=1}^{N} e^{-\frac{1}{2}(1-\xi_t^i)^2} P_{t-1}^*(i)}{\sum\limits_{i=1}^{N} e^{-\frac{1}{2}(1+\xi_t^i)^2} P_{t-1}^*(i)}\right).$$

Similar properties and assertions hold for the case where $\Omega = [-1, 1]$ (see Vovk [38]).

For infinite sets of outcomes Ω, the geometric definition of the mixable loss function has no meaning. In this case, we can introduce a general (direct) definition of mixability. The loss function is called η-mixable if there is a substitution function $\Sigma(g_t)$ such that

$$\lambda(\omega, \Sigma(g_t)) \leqslant g_y(\omega)$$

for all $\omega \in \Omega$, where g_t is defined by (5.18).

5.8. Universal portfolio selection

In this section we study the Cover's [9] game. Assume that there are N financial instruments (stocks) in Stock Market. The behavior of the market is specified by an arbitrary sequence of non-negative price relative stock vectors $\bar{\omega}_1, \bar{\omega}_2, \ \ldots\ ,$

$$\bar{\omega}_t = (\omega_{1,t}, \ \ldots, \ \omega_{N,t}), \tag{5.51}$$

The ith entry of tth price relative vector

$$\omega_{i,t} = \frac{S_{i,t+1}}{S_{i,t}}$$

denotes the ratio of closing $S_{i,t+1}$ to opening price $S_{i,t}$ of the ith stock for the tth trading day. By definition $\omega_{i,t} \in [0, \infty)$. We assume that $S_{i,t} > 0$ not all $\omega_{i,t}$ are zero.

The set of all allowable investments actions of algorithms we consider at round t is comprised of the state constant rebalanced portfolio that is a vector $\bar{\gamma}_t \in [0, 1]^N$, where

$$\bar{\gamma}_t = (\gamma_{1,t}, \ \ldots, \ \gamma_{N,t}) \tag{5.52}$$

and $\gamma_{1,t} + \cdots + \gamma_{N,t} = 1$.

The ith entry $\gamma_{i,t}$ of the portfolio is the proportion of wealth invested in the ith stock. An investment using a portfolio $\bar{\gamma}$ increases one's wealth by a factor of

$$(\bar{\gamma} \cdot \bar{\omega}) = \sum_{i=1}^{N} \gamma_i \omega_i$$

where the market performance is specified by the stock vector $\bar{\omega}$.

Further, for a sequence of T investments rounds, investing according to portfolios $\gamma_1, \ \ldots, \ \gamma_T$, increase the initial wealth by a factor of

$$S_T = \prod_{t=1}^{T} (\bar{\gamma}_t \cdot \bar{\omega}_t) = \prod_{t=1}^{T} \sum_{i=1}^{N} \gamma_{i,t} \omega_{i,t},$$

where a vector $\bar{\omega}_t = (\omega_{1,t}, \ \ldots, \ \omega_{N,t})$, characterizes the market performance at step t.

A sequence of portfolio choices $\bar{\gamma}_t$ constitutes an investment strategy of an algorithm.

We consider the problem of optimal investments in the prediction of expert advice framework. Define the corresponding loss function

$$\lambda(\bar{\omega}, \bar{\gamma}) = -\ln(\bar{\gamma} \cdot \bar{\omega}). \tag{5.53}$$

Here the set of outcomes Ω consists of all vectors $\bar{\omega}$ of the form (5.51) and the prediction set Γ is the simplex consisting of all vectors $\bar{\gamma}$ of the form (5.52) such that $\gamma_{1,t} + \cdots + \gamma_{N,t} = 1$.

A constant expert outputs a constant forecast – a portfolio $\bar{\gamma} \in \Gamma$. Conversely, any such constant portfolio defines an expert.

We apply the aggregating algorithm AA to this loss function following Vovk [37].

Assume that some a priory probability distribution $P_0(d\bar{\gamma})$ on the simplex Γ be given.

By (5.20) the pseudoprediction of APA is equal to

$$g_T(\omega) = \log_\beta \int_\Gamma \frac{\beta^{\lambda(\bar{\omega},\bar{\gamma}) + L_{T-1}(\bar{\gamma})}}{\int_\Gamma \beta^{L_{T-1}(\bar{\gamma})} P_0(d\bar{\gamma})} P_0(d\bar{\gamma}), \tag{5.54}$$

where $\beta = e^{-\eta}$, $0 < \eta \leqslant 1$.

Theorem 5.1. *The Cover's game (the loss function (5.53)) is η-mixable for any $0 < \eta \leqslant 1$. The corresponding substitution function is defined by the rule*

$$\Sigma(g_T) = \int_\Gamma \bar{\gamma} P_{T-1}(d\bar{\gamma}) =$$
$$= \int_\Theta \bar{\gamma} \frac{\beta^{L_{T-1}(\bar{\gamma})}}{\int_\Gamma \beta^{L_{T-1}(\gamma)} P_0(d\bar{\gamma})} P_0(d\bar{\gamma}). \tag{5.55}$$

Proof. We have to prove that for all $\bar{\omega}$

$$\lambda\left(\bar{\omega}, \int_\Gamma \bar{\gamma} P(d\bar{\gamma})\right) \leqslant \log_\beta \int_\Gamma \beta^{\lambda(\bar{\omega},\bar{\gamma})} P(d\bar{\gamma}).$$

This inequality is equivalent to the inequality

$$f\left(\int_\Gamma \bar{\gamma} P(d\bar{\gamma})\right) \geqslant \int_\Gamma f(\bar{\gamma}) P(d\bar{\gamma}), \tag{5.56}$$

where $f(\bar{\gamma}) = \beta^{\lambda(\bar{\omega},\bar{\gamma})} = (\bar{\gamma} \cdot \bar{\omega})^\eta$. The inequality (5.56) follows from concavity of the function $f(\bar{\gamma})$ for $0 < \eta \leqslant 1$. $\triangle$

We can write the portfolio (5.55) in more detail using the representation (5.21) of the cumulative loss for T rounds of the APA algorithm from Lemma 5.2:

$$L_T(APA) = \log_\beta \int_\Gamma \beta^{L_T(\bar{\gamma})} P_0(d\bar{\gamma}).$$

Put $\eta = 1$. Since the cumulative loss of any expert $\bar{\gamma}$ at first T steps is equal to

$$L_T(\bar{\gamma}) = -\ln \prod_{t=1}^{T} (\bar{\gamma} \cdot \bar{\omega}_t),$$

we can compute the forecast of the aggregating algorithm by the rule:

$$\bar{\gamma}_T = \frac{\int_\Gamma \bar{\gamma} \prod_{t=1}^{T-1} (\bar{\gamma} \cdot \bar{\omega}_t) P_0(d\bar{\gamma})}{\int_\Gamma \prod_{t=1}^{T-1} (\bar{\gamma} \cdot \bar{\omega}_t) P_0(d\bar{\gamma})}.$$

In general case it is convenient to consider the Dirichlet distribution with parameters $(1/2,\ \ldots,\ 1/2)$ on the simplex Γ:

$$P_0(d\bar{\gamma}) = \frac{\Gamma(N/2)}{[\Gamma(1/2)]^N} \prod_{j=1}^{N} \gamma_j^{-1/2} d\bar{\gamma},$$

where

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

Note that $\Gamma(N+1) = N!$.

We give without proof the main result of Cover's [9] which also was obtained by Vovk [36] as an application of the aggregating algorithm.

Theorem 5.2. *Let $\eta = 1$. Then the cumulative loss of the aggregating algorithm AA satisfies the inequality*

$$L_T(AA) \leqslant \inf_{\bar{\gamma}} L_T(\bar{\gamma}) + \frac{N-1}{2} \ln T + c \tag{5.57}$$

for all T, where c is a positive constant.

Since the total increase of the initial wealth under investment strategy presented by the aggregating algorithm is

$$K_T(AA) = e^{-L_T(AA)},$$

we can rewrite the inequality (5.57) in the form

$$K_T(AA) \geqslant T^{-\frac{N-1}{2}} K_T(\bar{\gamma}),$$

where $K_T(\bar{\gamma})$ is the total increase of the initial wealth using a constant rebalanced portfolio $\bar{\gamma}$.

From a theoretical perspective this is surprising as this performance ratio is bounded by a polynomial in T (for fixed N) whereas we suppose that the best portfolio in the growing market is capable of exponential returns. From a practical perspective, this bound is not very useful because the empirical returns of observed portfolios is often not exponential in the number of trading days.

5.9. Multidimensional online regression

In this section we consider the application of the aggregating algorithm AA to the problem of online regression. In contrast to conventional multidimensional regression which uses the training set to determine the parameters of regression once and for all, the AA-algorithm learns online.

Consider the problem of the online multidimensional linear regression in more detail.

Nature outputs pairs (x_t, y_t), where $x_t \in \mathcal{R}^n$ and $y_t \in \mathcal{R}$ for $t = 1, 2, \ldots$ [1] The regression problem consists in calculating at each step $t > 1$ a forecast for a future value y_t given the previously observed paits $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ and the input value x_t.

In the prediction with expert advice framework we introduce experts – linear functions $f(x) = (\theta \cdot x)$, where $\theta, x \in \mathcal{R}^n$. The values of these functions on inputs $x = x_t$ are interpreted as a forecast made by expert θ at step t. Denote by $\xi_t^\theta = (\theta \cdot x_t)$ this forecast.

The online regression problem with expert advice consists in computation at any step t a forecast γ_t of a future value y_t, using input x_t, past information, and predictions ξ_t^θ made by the linear experts at step t.

The general scheme of linear regression is regulated by the following perfect-information protocol of a game with players:. *Expert* θ, *Learner* and *Nature*.

FOR $t = 1, 2, \ldots$
Nature announces an input $x_t \in \mathcal{R}^n$.
Expert θ announces a forecast $\xi_t^\theta = (\theta \cdot x_t)$, $\theta \in \mathcal{R}^n$.
Learner announces a forecast $\gamma_t \in \mathcal{R}$.
announces an outcome $y_t \in [-Y, Y]$.
ENDFOR

A difference between the forecast and true value of the regression is measured by the square loss function. At any step t, the *Expert* θ loss is $(y_t - (\theta \cdot x_t))^2$, and the *Learner* loss is $(y_t - \gamma_t)^2$.

We apply the aggregating algorithm AA with a learning rate $\eta =$

[1] In this section we do not overbar vector variables.

$1/2Y^2$ to solve the problem of online regression.

The following a priory distribution on the set of experts will be used:

$$P_0(d\theta) = (a\eta/\pi)^{n/2} e^{-a\eta\|\theta\|^2} d\theta, \tag{5.58}$$

where a is a parameter (similar to a parameter used in the problem of ridge regression) and constants are chosen from the requirement of normalization. The Euclidian norm $\|\theta\| = \sqrt{\theta_1^2 + \cdots + \theta_n^2}$ is used for the vector $\theta = (\theta_1, \ \ldots, \ \theta_n)$.

Also, recall that we identify the dot product $(\theta \cdot x)$ and one-element matrix $x'\theta$, where x' is a row vector and θ is a column vector.

The loss of an expert $\theta \in \mathcal{R}^n$ at step t is equal to

$$\lambda(y_t, x_t'\theta) = (y_t - x_t'\theta)^2 = \theta'(x_t x_t')\theta - 2(y_t x_t')\theta + y_t^2. \tag{5.59}$$

Recall that $x_t' = (x_{1,t}, \ldots, \ x_{n,t})$, $\theta' = (\theta_1, \ldots, \ \theta_n)$ are row vectors and x_t, θ are column vectors. Here we also used the equality $x_t'\theta x_t'\theta = \theta'(x_t x_t')\theta$ that can be checked using the following transformations:

$$x_t'\theta x_t'\theta = \left(\sum_{i=1}^n x_{t,i}\theta_i\right)\left(\sum_{j=1}^n x_{t,j}\theta_j\right) = \\ = \sum_{i,j=1}^n \theta_i x_{t,i} x_{t,j}\theta_j = \theta'(x_t x_t')\theta.$$

The cumulative loss of the expert $\theta \in \mathcal{R}^n$ at first T steps is equal to

$$L_T(\theta) = \sum_{t=1}^T (y_t - x_t'\theta)^2 = \\ = \theta'\left(\sum_{t=1}^T x_t x_t'\right)\theta - 2\left(\sum_{t=1}^T y_t x_t'\right)\theta + \sum_{t=1}^T y_t^2. \tag{5.60}$$

By (5.8) and (5.22) we obtain

$$P_{t-1}(d\theta) = \beta^{L_{t-1}(\theta)} P_0(d\theta).$$

The pseudoprediction of the algorithm APA on step t is

$$g_t(y) = \log_\beta \int \beta^{\lambda(y,x_t'\theta)} P_{t-1}^*(d\theta) =$$

$$= \log_\beta \int \beta^{\lambda(y,x_t'\theta)} \frac{P_{t-1}(d\theta)}{P_{t-1}(\Theta)} =$$

$$= \log_\beta \int \beta^{\lambda(y,x_t'\theta)+L_{t-1}(\theta)} \frac{1}{P_{t-1}(\Theta)} P_0(d\theta). \tag{5.61}$$

Then, taking into account the representations (5.58) for the a priory distribution and (5.59) for the loss function, we obtain

$$g_T(-Y) = \log_\beta \int \beta^{\lambda(-Y,x_T'\theta)+L_{T-1}(\theta)} \frac{1}{P_{T-1}(\Theta)} P_0(d\theta) = \tag{5.62}$$

$$= \int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum_{t=1}^{T-1} y_t x_t' - Y x_T')\theta - \eta(\sum_{t=1}^{T-1} y_t^2 + Y^2)} \frac{d\theta}{P_{T-1}(\Theta)}.$$

A similar representation can be obtained for $g_T(Y)$.

We have proved in Section 5.7 that a substitution function exists for square loss function for the case of binary outcomes $\Omega = \{-1, 1\}$ and the prediction set $\Gamma = [-1, 1]$.

Similarly, it can be proved that the same substitution function is valid for the case where $\Omega = \Gamma = [-Y, Y]$ (see Vovk [38]).

Using the rule (5.62) for $g_T(-Y)$ and a similar rule for $g_T(Y)$, we

obtain

$$\gamma_T = \frac{1}{4Y}(g_T(-Y) - g_T(Y)) = \frac{1}{4Y} \times$$

$$\times \log_\beta \frac{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' - Y x_T')\theta - \eta(\sum\limits_{t=1}^{T-1} y_t^2 + Y^2)} d\theta}{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' + Y x_T')\theta - \eta(\sum\limits_{t=1}^{T-1} y_t^2 + Y^2)} d\theta} =$$

$$= \frac{1}{4Y} \log_\beta \frac{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' - Y x_T')\theta} d\theta}{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' + Y x_T')\theta} d\theta} =$$

$$= \frac{1}{4Y} \log_\beta e^{-\eta F\left(aI+\sum\limits_{t=1}^{T} x_t x_t', -2\sum\limits_{t=1}^{T-1} y_t x_t', 2Y x_T'\right)} =$$

$$= \frac{1}{4Y} F\left(aI + \sum_{t=1}^{T} x_t x_t', -2\sum_{t=1}^{T-1} y_t x_t', 2Y x_T'\right) =$$

$$= \left(\sum_{t=1}^{T-1} y_t x_t'\right)\left(aI + \sum_{t=1}^{T} x_t x_t'\right)^{-1} \cdot x_T. \qquad (5.63)$$

Here we at once reduced the common factor $\frac{1}{P_{T-1}(\Theta)}$ in the numerator and denominator of the 2th row. In transition from 2th row to the 3th, the factor $e^{-\eta(\sum\limits_{t=1}^{T-1} y_t^2 + Y^2)}$ was taken out from the integral in the numerator and denominator and reduced. In transition from 3th to 4th we have used Lemma 5.4 given below which says that the integral in the numerator of the 3th row is equal to

$$\frac{\pi^{n/2}}{\sqrt{\det A}} e^{-\eta \inf\limits_{\theta\in\mathcal{R}^n}(\theta' A\theta + c'\theta + x'\theta)},$$

and the integral in the denominator of the 3th row is equal to

$$\frac{\pi^{n/2}}{\sqrt{\det A}} e^{-\eta \inf\limits_{\theta\in\mathcal{R}^n}(\theta' A\theta + c'\theta - x'\theta)},$$

where

$$A = aI + \sum_{t=1}^{T} x_t x_t',$$
$$c = -2 \sum_{t=1}^{T-1} y_t x_t',$$
$$x = 2Y x_T'.$$

In the 4th row we have used notation

$$F(A,c,x) = \inf_{\theta \in \mathcal{R}^n} (\theta' A\theta + c'\theta + x'\theta) - \\ - \inf_{\theta \in \mathcal{R}^n} (\theta' A\theta + c'\theta - x'\theta), \tag{5.64}$$

and in transition from the 5th row to the 6th row we have used Lemma 5.5 which says that $F(A,c,x) = -c'A^{-1}x$.

Now we give a formulation and proofs of Lemmas 5.4 and 5.5.

Lemma 5.4. *Let $Q(\theta) = \theta' A\theta + c'\theta + d$, where $\theta, c \in \mathcal{R}^n$, $d \in \mathcal{R}$ and A be symmetric positive definite matrix of type $(n \times n)$. Then*

$$\int_{\mathcal{R}^n} e^{-Q(\theta)} d\theta = e^{-Q_0} \frac{\pi^{n/2}}{\sqrt{\det A}}, \tag{5.65}$$

where $Q_0 = \min_{\theta \in \mathcal{R}^n} Q(\theta)$.

Proof. Assume that a minimum of the quadratic form

$$Q(\theta) = \theta' A\theta + c'\theta + d$$

is attained at $\theta = \theta_0$. Denote $\xi = \theta - \theta_0$ and $\tilde{Q}(\xi) = Q(\xi + \theta_0)$. It is easy to see that the quadratic part of the form $\tilde{Q}$ is $\xi' A\xi$. since the minimum of the quadratic form $\tilde{Q}$ is attained at $\theta = \bar{0}$, where $\bar{0} = (0, \ \ldots, \ 0)$, this form cannot have a linear part. Indeed, otherwise, this linear part should dominate under the quadratic part of this form in a small neighborhood of $\bar{0}$, and then the form $\tilde{Q}$ could not attain a minimum at $\bar{0}$.

Since the minimum of the form $\tilde{Q}(\xi)$ is equal to Q_0, the constant of this form is Q_0. Therefore, $\tilde{Q}(\xi) = \xi' A \xi + Q_0$.

It remains to prove that

$$\int\limits_{\mathcal{R}^n} e^{-\xi' A \xi} d\xi = \pi^{n/2} / \sqrt{\det A}.$$

This follows from Theorem 3 (Section 2.7) of Bellman [5]. $\triangle$

Lemma 5.5 below shows that $F(A, c, x) = -c' A^{-1} x$.

Lemma 5.5. *Let A be a symmetric positive definite matrix of type $(n \times n)$ and $b, x \in \mathcal{R}^n$. Then*

$$F(A, c, x) = \min_{\theta \in \mathcal{R}^n} (\theta' A \theta + c' \theta + x' \theta) -$$
$$- \min_{\theta \in \mathcal{R}^n} (\theta' A \theta + c' \theta - x' \theta) = -c' A^{-1} x. \tag{5.66}$$

Proof. To find the first minimum equate to zero the partial derivatives of the quadratic form $\theta' A \theta + c' \theta + x' \theta$ by θ_i. Here $\theta = (\theta_1, \ \ldots, \ \theta_n)$. As a result we obtain a system of equations $2A\theta + c' + x' = \bar{0}$. It is easy to see that the minimum is attained at $\theta_1 = -\frac{1}{2} A^{-1}(c + x)$. Similarly, the minimum of the second part $\theta_1 = -\frac{1}{2} A^{-1}(c - x)$. After that, the assertion of the lemma is obtained by substituting these values into the difference of these terms. $\triangle$

Therefore, according to the expression (5.63) for γ_T, at step T, we have

$$A = aI + \sum_{t=1}^{T} x_t x_t',$$
$$b = \sum_{t=1}^{T-1} y_t x_t',$$
$$\gamma_T = b' A^{-1} x_T =$$
$$= \left(\sum_{t=1}^{T-1} y_t x_t' \right) \left(aI + \sum_{t=1}^{T} x_t x_t' \right)^{-1} \cdot x_T.$$

Now we can rewrite the regression algorithm AAR in the form:

$A = aI$; $b' = \bar{0}$.
FOR $t = 1, 2, \ \ldots$
Algorithm receives an input $x_t \in \mathcal{R}^n$.
Compute $A = A + x_t x_t'$.
Output a forecast $\gamma_t = b' A^{-1} x_t$.
Algorithm receives $y_t \in [-Y, Y]$.
Compute $b' = b' + y_t x_t'$.
ENDFOR

Comparing the cumulative loss of the algorithm AAR and the cumulative loss of the best expert, we obtain the following result.

Theorem 5.3. *For any T,*

$$L_T(AAR) \leqslant \inf_\theta (L_T(\theta) + a\|\theta\|_2^2) + Y^2 \ln \det \left(I + \frac{1}{a} \sum_{t=1}^T x_t x_t' \right) \leqslant$$

$$\leqslant \inf_\theta (L_T(\theta) + a\|\theta\|_2^2) + Y^2 \sum_{i=1}^n \ln \left(1 + \frac{1}{a} \sum_{t=1}^T x_{t,i}^2 \right).$$

If in addition $|x_{t,i}| \leqslant X$ for all t and i then

$$L_T(AAR) \leqslant \inf_\theta (L_T(\theta) + a\|\theta\|_2^2) + nY^2 \ln \left(\frac{TX^2}{a} + 1 \right).$$

Proof. Let $\eta = \frac{1}{2Y^2}$. By Lemma 5.2 the cumulative loss of the algorithm APA is represented in the form (5.21). In this case we can rewrite this expression as

$$L_T(APA) = \log_\beta \int_{\mathcal{R}^n} e^{-\eta L_T(\theta)} P_0(d\theta) =$$

$$= \log_\beta \int_{\mathcal{R}^n} (a\eta/\pi)^{n/2} e^{-\eta\theta'(aI + \sum\limits_{t=1}^T x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^T y_t x_t')\theta - \eta \sum\limits_{t=1}^T y_t^2} d\theta. \quad (5.67)$$

The exponent in (5.67) has a form $e^{-\eta F(\theta)}$, where

$$F(\theta) = \theta' \left(aI + \sum_{t=1}^T x_t x_t' \right) \theta - 2 \left(\sum_{t=1}^T y_t x_t' \right) \theta + \sum_{t=1}^T y_t^2.$$

Assume that the minimum of $F(\theta)$ is attained at $\theta = \theta_0$.

Then by Lemma 5.4 the expression (5.67) representing the cumulative loss of the algorithm APA is equal to

$$L_T(APA) =$$

$$= \log_\beta \left(((a\eta/\pi)^{n/2}) \frac{\pi^{n/2} e^{-\eta F(\theta_0)}}{\sqrt{\det\left(a\eta I + \eta \sum_{t=1}^{T} x_t x_t'\right)}} \right) =$$

$$= F(\theta_0) - \frac{1}{2} \log_\beta \det\left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t'\right) =$$

$$= F(\theta_0) + \frac{1}{2\eta} \ln \det\left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t'\right) =$$

$$= F(\theta_0) + Y^2 \ln \det\left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t'\right).$$

By definition (5.60) of cumulative loss of an expert θ

$$F(\theta_0) = \theta_0' \left(aI + \sum_{t=1}^{T} x_t x_t'\right) - 2\left(\sum_{t=1}^{T} y_t x_t'\right)\theta_0 + \sum_{t=1}^{T} y_t^2 =$$

$$= a\|\theta_0\|^2 + \sum_{t=1}^{T} (y_t - x_t'\theta_0)^2 =$$

$$= a\|\theta_0\|^2 + L_T(\theta_0).$$

Since $L_T(AAR) \leqslant L_T(APA)$, we have the assertion ot the theorem. $\triangle$

5.10. Multidimensional kernel regression

To move from linear to kernel regression we have to translate all algorithms in a form where they depend only from the inner products of the input variables. We assume that a regression hyperplane was

carried out in the feature space. After that, all inner products will be represented by the kernel values.

Recall the algorithm of linear regression:

$$A = aI + \sum_{t=1}^{T} x_t x_t',$$

$$b = \sum_{t=1}^{T-1} y_t x_t',$$

$$\gamma_T = b' A^{-1} x_T =$$

$$= \left(\sum_{t=1}^{T-1} y_t x_t' \right) \left(aI + \sum_{t=1}^{T} x_t x_t' \right)^{-1} \cdot x_T. \qquad (5.68)$$

Let $K(x, x')$ be a kernel, where $x, x' \in \mathcal{R}^n$, and $S = ((\bar{x}_1, y_1), (\bar{x}_2, y_2), \dots)$ be an unput sample.

Denote:

$K_T = (K(x_i, x_j))_{i,j=1}^{T}$ – kernel values matrix;

$k_T = (k(x_i, x_T))_{i=1}^{T}$ – the last column of the matrix K_T;

Y_T – a column-vector of outcomes;

$(Y_{T-1}, 0) = (y_1, \dots, y_{T-1}, 0)$ – an incomplete column vector complemented by zero.

Let us present the linear regression algorithm (5.68) in the appropriate form to obtain its kernel version.

Consider the matrix of type $T \times n$:

$$X_T = \begin{pmatrix} x_1' \\ x_2' \\ \dots \\ x_T' \end{pmatrix} = \begin{pmatrix} x_{11}, x_{12}, \dots, x_{1T} \\ x_{21}, x_{22}, \dots, x_{2T} \\ \dots \\ x_{n1}, x_{n2}, \dots, x_{nT} \end{pmatrix},$$

in which the rows are row-vectors $x_1', \dots, x_T'$.

It is easy to verify that

$$\sum_{t=1}^{T} x_t x_t' = X_T' X_T,$$

and also,

$$\sum_{t=1}^{T-1} y_t x_t = (Y_{T-1}, 0)' X_T.$$

The following lemma holds:

Lemma 5.6. *For any matrix B of type $n \times m$ and any matrix C of type $m \times n$ such that the matrixes $aI_n + CB$ and $aI_m + BC$ are invertible,*

$$B(aI_n + CB)^{-1} = (aI_m + BC)^{-1} B, \tag{5.69}$$

where a is a real number and I_n is the unit matrix of size n. [2]

Proof. The equality (5.69) is equivalent to the equality

$$(aI_n + BC)B = B(aI_m + CB),$$

that is evident, since the matrix product is distributive. $\triangle$

Using this lemma, we present the prediction of the linear regression as follows:

$$\begin{aligned} \gamma_T = b' A^{-1} x_T = \\ = \left(\sum_{t=1}^{T-1} y_t x_t' \right) \left(aI + \sum_{t=1}^{T} x_t x_t' \right)^{-1} \cdot x_T = \\ = (Y_{T-1}, 0)' \, X_T \left(aI + X_T' X_T \right)^{-1} x_T = \\ (Y_{T-1}, 0)' \left(aI + X_T X_T' \right)^{-1} X_T x_T = \\ = (Y_{T-1}, 0)' \left(aI + \tilde{K}_T \right)^{-1} \tilde{k}_T, \end{aligned} \tag{5.70}$$

where $\tilde{K}_T = X_T X_T'$ and $\tilde{k}_T = X_T x_T$. Note also, that

$$\begin{aligned} X_T X_T' = (x_t \cdot x_{t'})_{t,t'=1}^{T}, \\ \tilde{k}_T = (x_t \cdot x_T)_{t=1}^{T}, \end{aligned}$$

[2] In what follows the index n is omitted.

ie, the coordinates of the matrix and of the vector are dot product of vectors $x_1, \ldots, x_T$.

We obtain the adaptive algorithm of the kernel version by replacing the dot products from the matrix $\tilde{K}_T = X_T X_T'$ and the vector $\tilde{k}_T = X_T x_T$ of the linear version by the corresponding kernel values $K_T = (K(x_i, x_j))_{i,j=1}^T$ and $k_T = (k(x_i, x_T))_{i=1}^T$. As a result, we obtain a forecast of the kernel version

$$\gamma_T = (Y_{T-1}, 0)' (aI + K_T)^{-1} k_T.$$

A bound for the learning error of the kernel regression has a form

Theorem 5.4. *For all T*

$$L_T(AAR) \leqslant \inf_\theta (L_T(\theta) + a|\theta|_2^2) + Y^2 \ln det \left(\frac{\tilde{K}_T}{a} + I \right)$$

for all T.

Dual form of the kernel regression

We now give a definition of the dual form of any prediction algorithm. Assume that a forecasting algorithm $\mathcal{A}$ uses input vectors $x_1, x_2, \ldots x_T$ only in the form of dot products and a kernel function $K(x, y)$, where $x, y \in \mathcal{R}^n$, be given.

Using Lemma 5.6, we transform the forecast

$$\gamma_{T+1} = w'x = ((aI + X_T' X_T)^{-1} X_T' Y_T)' \cdot x_{T+1}$$

of the ridge regression to the form

$$\begin{aligned} \gamma_{T+1} = w'x = ((aI + X_T' X_T)^{-1} X_T' Y_T)' \cdot x_{T+1} = \\ = Y_T' X_T (aI + X_T' X_T)^{-1} \cdot x_{T+1} = \\ = Y_T' (aI + X_T X_T')^{-1} X_T \cdot x_{T+1} = \\ = Y_T' (aI + X_T X_T')^{-1} X_T \cdot k_{T+1}. \end{aligned}$$

Note that in the same notation, the expression (5.70) for adaptive kernel regression has a slightly different form:

$$\gamma_{T+1} = (Y_T, 0)' (aI + K_{T+1})^{-1} \cdot k_{T+1}.$$

5.11. Laboratory work

Use the data from the following websites for solving problems of regression:

$$http://www.csie.ntu.edu.tw$$

Database UCI Machine Learning Repository is located on the website

$$http://archive.ics.uci.edu$$

Laboratory work 1

Construct a simple linear, ridge regression, and regression with standard SVM software for data from UCI repository. Provide a comparative analysis of the accuracy of regression for all methods used.

Laboratory work 2

Conduct also experiments with kernel versions of these methods. Provide a comparative analysis of the accuracy of regression for all methods used.

Laboratory work 3

Conduct a linear online regression using aggregating algorithm from Section 5.9. Compare the accuracy of the regression with other methods.

5.12. Problems

1. Draw the graphs of the prediction and superprediction sets and of their images in the exponential space for the square, logarithmic, absolute and simple loss functions for different values of η. Give examples of η where the corresponding images of the prediction and superprediction sets for the logarithmic and square loss functions are convex and non-convex.

2. Prove that the absolute loss function $\lambda(\omega, \gamma) = |\omega - \gamma|$ is not mixable. Calculate the multiplicative constants $c(\beta)$ and $a(\beta)$ from the bound (5.27) for an arbitrary $\beta = e^{-\eta}$, $\eta > 0$, in the case of

absolute loss function and $\omega \in \{0,1\}$, $\gamma \in [0,1]$. (*Hint*:

$$c(\eta) = \frac{\frac{1}{2}\eta}{\ln\left(\frac{2}{1+e^{-\eta}}\right)} = \frac{\frac{1}{2}\ln\frac{1}{\beta}}{\ln\left(\frac{2}{1+\beta}\right)},$$

where $\beta = e^{-\eta}$.

This expression can be obtained by minimizing the term $c(\eta)$ in the exponent:

$$e^{-\eta\lambda(\omega,\gamma} \geqslant \left(\sum_{i=1}^{N} e^{-\eta\lambda(\omega,\xi_i)}\right)^{c(\eta)}$$

for all ω, γ and ξ_i. It is convenient to use the geometric representation – the curve:

$$\left(e^{-\eta\gamma}, e^{-\eta(1-\gamma)}\right)$$

where $0 \leqslant \gamma \leqslant 1$. Find $c(\eta)$ for which the distance from the point on the curve to the chord of maximal length is maximal for all ω, γ and ξ_i).

Find an expression for the prediction of the aggregating algorithm in this case.

3. Show that the curve (5.37) is convex.

4. Draw the graph of the curve $c(\beta)$ defined by the equality (5.38).

5. Proof the inequality (5.39). Study the dependence of the factor in the inequality of (5.39) from the parameter β.

Rebuild this estimate in an estimate with the unit factor and regretom of order $O(\sqrt{T\ln N})$ in the same way as has been done in the the allocation algorithm $Hedge(\beta)$ Section 4.2.

Part III

Games of Prediction

Chapter 6

Elements of the game theory

In this chapter, we first consider the classical problems of the game theory, namely, two-person zero-sum games. We will prove the von Neumann minimax theorem and consider methods of solving such games. Further, in Chapter 8, we apply the minimax theorem to solve infinitely repeatable games of prediction.

6.1. Two players zero-sum games

Let X and Y be arbitrary sets. Let us consider two-person zero-sum game. The first player chooses an action (or a strategy) $x \in X$; simultaneously with this the second player chooses an action $y \in Y$.

When a game is presented in normal form, it is presumed that each player acts simultaneously or, at least, without knowing the actions of the other player. As a result of their actions players receive payoff or suffer loss. More generally, this payoff or loss can be represented by any function that associates a payoff for each player with every possible combination of actions.

A function $f(x, y)$ represents the payoff for first player, which is also a function of the loss of the second player. The function $f(x, y)$ is defined on the Cartesian product $X \times Y$.

If $f(x, y) < 0$, the payoff for the first player is negative, ie, this player suffers a loss.

Each player is supposed to behave rationally. This means that each player tries to maximize his payoff irrespective to what other player is doing.

In zero sum game, sum of payoffs of all the players for each round of the game is zero. Which means if one player is able to improve his payoff by using some good strategy the payoff of others is going to decrease. In zero-sum game, the goal of the first player is to maximize his payoff and the goal of the second player is to minimize his loss.

If the first player chooses a strategy x, then its payoff is not less than than $\inf_{y \in Y} f(x, y)$ independently of the choice of the second player. This value is called the guaranteed result for the first player. The best guaranteed result for the first player

$$\underline{v} = \sup_{x \in X} \inf_{y \in Y} f(x, y)$$

is called *the lower value* of the game.

A strategy of x^0 of the first player is called *maxmin* if

$$\inf_{y \in Y} f(x^0, y) = \underline{v}.$$

In terms of the second player, the choice of strategy y guarantees him a loss no more than $\sup_{x \in X} f(x, y)$ – his *a guaranteed result.* The best guaranteed result of the second player

$$\overline{v} = \inf_{y \in Y} \sup_{x \in X} f(x, y)$$

is called *the upper value* of the game.

A strategy y^0 of the second player is called minimax strategy if

$$\sup_{x \in X} f(x, y^0) = \overline{v}.$$

Lemma 6.1. *In any zero-sum game* $\underline{v} \leqslant \overline{v}$*, ie,*

$$\sup_{x \in X} \inf_{y \in Y} f(x, y) \leqslant \inf_{y \in Y} \sup_{x \in X} f(x, y).$$

Proof. For any $x \in X$ and $y \in Y$,

$$\inf_{y\in Y} f(x,y) \leqslant f(x,y) \leqslant \sup_{x\in X} f(x,y).$$

Then

$$\inf_{y\in Y} f(x,y) \leqslant \sup_{x\in X} f(x,y).$$

The left-hand side of this inequality depends on x, and the right-hand side does not depend of x. Then

$$\sup_{x\in X} \inf_{y\in Y} f(x,y) \leqslant \sup_{x\in X} f(x,y)$$

for all y. Hence,

$$\underline{v} = \sup_{x\in X} \inf_{y\in Y} f(x,y) \leqslant \inf_{y\in Y} \sup_{x\in X} f(x,y) = \overline{v}.$$

Lemma is proved. $\triangle$

A point $(x^0, y^0) \in X \times Y$ is called *saddle point* of f if

$$f(x,y^0) \leqslant f(x^0,y^0) \leqslant f(x^0,y) \tag{6.1}$$

for all $x \in X$ and $y \in Y$.

The condition (6.1) is equivalent to

$$\max_{x\in X} f(x,y^0) = f(x^0,y^0) = \min_{y\in Y} f(x^0,y). \tag{6.2}$$

Note that, when we write min instead of inf or max instead of sup, then we have in mind that these extreme values are attained at some points.

We say that a zero-sum game has a solution if the function $f(x,y)$ has a saddle point (x^0, y^0). The number $v = f(x^0, y^0)$ is called *the value* or *price* of the game, x^0, y^0 are optimal strategies of the players. The number $v = f(x^0, y^0)$ is called *solution* of the game. These titles are justified by the following theorem.

Theorem 6.1. *1. A function $f(x,y)$ has a saddle point if and only if*

$$\max_{x\in X} \inf_{y\in Y} f(x,y) = \min_{y\in Y} \sup_{x\in X} f(x,y). \tag{6.3}$$

2. *Let (6.3) holds. Then a pair (x^0, y^0) is a saddle point if and only if x^0 is the maxmin strategy and y^0 is the minimax strategy.*

Proof of necessity of 1) and 2). Let (x^0, y^0) be a saddle point of the function $f(x, y)$. Then

$$\overline{v} \leqslant \sup_{x \in X} f(x, y^0) = f(x^0, y^0) = v = \inf_{y \in Y} f(x^0, y) \leqslant \underline{v}. \tag{6.4}$$

From this $\overline{v} \leqslant \underline{v}$. By Lemma 6.1, the equality $\overline{v} = \underline{v}$ holds. Then (6.4) is also the equality, and then x^0 is the maxmin strategy and y^0 is the minimax strategy.

Proof of sufficiency. Assume that (6.3) is valid. Let x^0 be a maxmin and y^0 be a minimax strategy. We show that (x_0, y_0) is a saddle point. Indeed,

$$f(x^0, y^0) \geqslant \inf_{y \in Y} f(x^0, y) = \underline{v} = \overline{v} = \sup_{x \in X} f(x, y^0) \geqslant f(x^0, y^0).$$

This implies that all these inequalities are equalities. Therefore, (x^0, y^0) is a saddle point. $\triangle$

Example. Game of matching pennies is played between two players, player A and player B. Each player has a penny and must secretly turn the penny to heads or tails. The players then reveal their choices simultaneously. If the pennies match (both heads or both tails) player A keeps both pennies, so wins one from player B (+1 for A and −1 for B). If the pennies do not match (one heads and one tails) player B keeps both pennies, so receives one from player A (−1 for A and +1 for B). The game can be written in a payoff matrix

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

Evidently, this game has not a saddle point. The best guaranteed result for the first player is $\underline{v} = \max_i \min_j a_{i,j} = -1$, and the best guaranteed result for the second player is $\overline{v} = \min_j \max_i a_{i,j} = 1$. This game does not have a solution.

6.2. Sufficient condition for the existence of a saddle point

We give a sufficient condition for the existence of a saddle point, which implies the minimax theorem.

At first recall that a subset $Z \subseteq \mathcal{R}^n$ of the Euclidian space $\mathcal{R}^n$ is called convex if for any two points $z, z' \in Z$ and for any real number $0 \leqslant p \leqslant 1$ the point $pz + (1-p)z' \in Z$.

A function $h(z)$ defined on a convex set Z is called convex if for any $z, z' \in Z$ and for any real number $0 \leqslant p \leqslant 1$ the following inequality

$$h(pz + (1-p)z') \leqslant ph(z) + (1-p)h(z') \tag{6.5}$$

holds.

The function $h(z)$ is called concave if the inequality (6.5) holds, where the symbol $\leqslant$ is replaced by $\geqslant$.

Theorem 6.2. *Let X and Y be convex subsets of $\mathcal{R}^n$ and $\mathcal{R}^m$ respectively, where n and m be arbitrary positive integer numbers, Y be a compact set. Let also,*

- *the real function $f(x, y)$ be defined on $X \times Y$ and bounded in absolute value,*
- *the function $f(x, \cdot)$ be convex and continuous by y for each value of $x \in X$,*
- *$f(\cdot, y)$ be concave for each value of $y \in Y$.*

Then

$$\sup_{x \in X} \inf_{y \in Y} f(x, y) = \inf_{y \in Y} \sup_{x \in X} f(x, y).$$

Proof. By Lemma 6.1 we have to prove that

$$\inf_{y \in Y} \sup_{x \in X} f(x, y) \leqslant \sup_{x \in X} \inf_{y \in Y} f(x, y).$$

Assume without loss of generality that $f(x, y) \in [0, 1]$.

Fix a sufficiently small real number $\epsilon > 0$ and a sufficiently large positive integer number n. By compactness of Y an ϵ-net $\{y^1, \ \ldots, \ y^N\}$ exists such that any point $y \in Y$ is located in the ϵ-neighborhood of some point y^i.

Define a sequence of points $y_1, y_2, \ \ldots, \ y_n \in Y$ and a sequence of points $x_1, x_2, \ \ldots, \ x_n \in X$ recursively. Let x_0 be any point of X. Define for $t = 1, \ \ldots, \ n$:

$$y_t = \frac{\sum\limits_{i=1}^{N} y^i e^{-\eta \sum_{s=0}^{t-1} f(x_s, y^i)}}{\sum\limits_{j=1}^{N} e^{-\eta \sum_{s=0}^{t-1} f(x_s, y^j)}}, \tag{6.6}$$

where $\eta = \sqrt{(8 \ln N)/n}$ and a point x_t is defined such that

$$f(x_t, y_t) \geqslant \sup_{x \in X} f(x, y_t) - \frac{1}{n}.$$

Since the function f is convex by the second argument, we can use Theorem 4.6, where the loss function is $\lambda(x, y) = f(x, y)$.

Let in the exponentially weighted forecaster (6.6) y^i be the expert forecasts, $i = 1, \ldots, N$, x_t be outcomes, $t = 1, \ldots, n$, and y_t be *Learner* forecast. By (4.44) we have

$$\sum_{t=1}^{n} f(x_t, y_t) \leqslant \min_{i=1,\ldots,N} \sum_{t=1}^{n} f(x_t, y^i) + \sqrt{\frac{1}{2} n \ln N}.$$

We divide this inequality by n :

$$\frac{1}{n} \sum_{t=1}^{n} f(x_t, y_t) \leqslant \min_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} f(x_t, y^i) + \sqrt{\frac{\ln N}{2n}}. \tag{6.7}$$

Since the function f is convex by the second argument and is concave

by the first argument and (6.7) holds, we have

$$\begin{aligned}
&\inf_{y\in Y}\sup_{x\in X} f(x,y) \leqslant \\
&\leqslant \sup_{x\in X} f\left(x, \frac{1}{n}\sum_{t=1}^{n} y_t\right) \leqslant \\
&\leqslant \sup_{x\in X} \frac{1}{n}\sum_{t=1}^{n} f(x,y_t) \leqslant \\
&\leqslant \frac{1}{n}\sum_{t=1}^{n} \sup_{x\in X} f(x,y_t) \leqslant \\
&\leqslant \frac{1}{n}\sum_{t=1}^{n} f(x_t,y_t) + \frac{1}{n} \leqslant \\
&\leqslant \min_{i=1,\ ...,\ N} \frac{1}{n}\sum_{t=1}^{n} f(x_t,y^i) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n} \leqslant \\
&\leqslant \min_{i=1,\ ...,\ N} f\left(\frac{1}{n}\sum_{t=1}^{n} x_t, y^i\right) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n} \leqslant \\
&\leqslant \sup_{x\in X} \min_{i=1,\ ...,\ N} f(x,y^i) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n}. \qquad (6.8)
\end{aligned}$$

The transition from the 1st line to the 2nd is by definition; the transition from 2nd to 3rd line is by convexity of $f(x,\cdot)$; the transition from 3rd 4th line is made because the supremum of the sum does not exceed the sum of suprema; the transition from the 4th to the 5th line is by definition of x_t; the transition from 5th to 6th line is by (6.7); the transition from 6th to 7th line is by concavity of the function $f(\cdot,y)$; the transition from 7th to 8th line is by by definition of the supremum.

Therefore, we have proved that for all n

$$\inf_{y\in Y}\sup_{x\in X} f(x,y) \leqslant \sup_{x\in X} \min_{i=1,\ ...,\ N} f(x,y^i) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n}.$$

Tending n to infinity, we obtain

$$\inf_{y\in Y}\sup_{x\in X} f(x,y) \leqslant \sup_{x\in X} \min_{i=1,\ldots,N} f(x,y^i).$$

Tending $\epsilon \to 0$, we obtain

$$\inf_{y\in Y}\sup_{x\in X} f(x,y) \leqslant \sup_{x\in X}\inf_{y\in Y} f(x,y).$$

Theorem is proved. $\triangle$

The proof of Theorem 6.2 contains a method for calculating the value of the game, since from the 1st, 5th and 8th lines of inequality (6.8) follows that the value of $\frac{1}{n}\sum\limits_{t=1}^{n} f(x_t, y_t)$ is arbitrarily close approximation to the value of the game for a sufficiently small ϵ and sufficiently large n.

6.3. Mixed extension of matrix games

6.3.1. Minimax theorem

Assume that $X = \{1, \ \ldots, \ N\}$ and $Y = \{1, \ \ldots, \ M\}$ be sets of strategies of the first and the second player. The corresponding game is called matrix game, since the payoff function $f(i,j) = a_{i,j}$ can be represented as a matrix. The first player chooses a row, the second player chooses a number. The element $a_{i,j}$ located in their intersection determines the gain of the first player that is the loss the second player.

Now let players make their choices at random. *A mixed strategy* of a player is a probability distribution on set of of his moves. Mixed extension of the matrix game $(X, Y, f(x,y))$ is defined as a game $(\mathcal{X}, \mathcal{Y}, \bar{f}(\bar{p}, \bar{q}))$, where $\mathcal{X}$ is the set of all mixed strategies of the first player and $\mathcal{Y}$ is the set of all mixed strategies of the second player, $\bar{f}(\bar{p}, \bar{q})$ is the mathematical expectation of the payoff function $f(i,j)$

with respect to the probability distribution $p \times q$:

$$\mathcal{X} = \{\bar{p} = (p_1, \ \ldots, \ p_N) : \sum_{i=1}^{N} p_i = 1, \ p_i \geqslant 0\};$$

$$\mathcal{Y} = \{\bar{q} = (q_1, \ \ldots, \ q_M) : \sum_{i=1}^{M} q_i = 1, \ q_i \geqslant 0\};$$

$$\bar{f}(\bar{p}, \bar{q}) = \sum_{i=1}^{N} \sum_{j=1}^{M} f(i,j) p_i q_j.$$

Takes place minimax theorem of von Neumann.

Theorem 6.3. *Any matrix game has a solution in mixed strategies:*

$$\max_{\bar{p} \in \mathcal{X}} \min_{\bar{q} \in \mathcal{Y}} \bar{f}(\bar{p}, \bar{q}) = \min_{\bar{q} \in \mathcal{Y}} \max_{\bar{p} \in \mathcal{X}} \bar{f}(\bar{p}, \bar{q}).$$

Proof. It suffices to prove that the function $\bar{f}(\bar{p}, \bar{q})$ has a saddle point. We shall apply Theorem 6.2. The sets $\mathcal{X}$ and $\mathcal{Y}$ are simplexes in Euclidean space, and so, they are convex. The function $\bar{f}(\bar{p}, \bar{q})$ is bilinear and therefore continuous by both arguments, concave and convex by them. $\triangle$

6.3.2. Pure strategies

Consider a matrix game with sets of strategies $X = \{1, \ \ldots, \ N\}$, $Y = \{1, \ \ldots, \ M\}$ and with a payoff function $f(i,j) = a_{i,j}$. We give three simple statements that describe in more detail the structure of the optimal solution in terms of pure strategies.

Denote by $1_i = (0, \ \ldots, \ 1, \ldots, \ 0)$ the pure strategy that is a probability distribution concentrated in $i \in X$. This is the unit vector of length N whose ith coordinate is 1. Similarly, we define pure strategies on the set Y. Notice that $\bar{f}(1_i, 1_j) = f(i,j) = a_{i,j}$.

Theorem 6.4. *The pair of mixed strategies $(\bar{p}^*, \bar{q}^*)$ is a solution or a saddle point of the mixed extension of the matrix game $(\mathcal{X}, \mathcal{Y}, \bar{f}(\bar{p}, \bar{q}))$ if and only if the inequality*

$$\bar{f}(1_i, \bar{q}^*) \leqslant \bar{f}(\bar{p}^*, \bar{q}^*) \leqslant \bar{f}(\bar{p}^*, 1_j) \tag{6.9}$$

holds for all $i \in X$ and $j \in Y$.

Proof. The necessity follows from Theorem 6.1. To prove the sufficient condition note that every mixed strategy $\bar{p} = (p_1, \ \ldots, \ p_N)$ of the matrix game is a linear combination of the pure strategies: $\bar{p} = \sum\limits_{i=1}^{N} p_i 1_i$. Similarly, $\bar{q} = \sum\limits_{j=1}^{M} q_j 1_j$. Therefore, we can consider the double linear combination of the inequality (6.9). Then

$$\bar{f}(\bar{p}, \bar{q}^*) = \sum_{i=1}^{N} p_i \bar{f}(1_i, \bar{q}^*) \leqslant \sum_{i=1}^{N} p_i \bar{f}(\bar{p}^*, \bar{q}^*) = \bar{f}(\bar{p}^*, \bar{q}^*),$$
$$\bar{f}(\bar{p}^*, \bar{q}^*) = \sum_{j=1}^{M} q_j \bar{f}(\bar{p}^*, \bar{q}^*) \leqslant \sum_{j=1}^{M} q_j \bar{f}(\bar{p}^*, 1_j) = \bar{f}(\bar{p}^*, \bar{q})$$

for all $\bar{p}$ and $\bar{q}$. Hence we obtain the saddle point condition:

$$\bar{f}(\bar{p}, \bar{q}^*) \leqslant \bar{f}(\bar{p}^*, \bar{q}^*) \leqslant \bar{f}(\bar{p}^*, \bar{q})$$

for all $\bar{p}$ and $\bar{q}$. $\triangle$

Theorem 6.5. *For the mixed extension of any matrix game the following relations hold:*

$$\min_{\bar{q}} \bar{f}(\bar{p}, \bar{q}) = \min_{j} \bar{f}(\bar{p}, 1_j),$$
$$\max_{\bar{p}} \bar{f}(\bar{p}, \bar{q}) = \max_{i} \bar{f}(1_i, \bar{q}).$$

Proof. Evidently,

$$\min_{\bar{q}} \bar{f}(\bar{p}, \bar{q}) \leqslant \min_{j} \bar{f}(\bar{p}, 1_j),$$
$$\max_{\bar{p}} \bar{f}(\bar{p}, \bar{q}) \geqslant \max_{i} \bar{f}(1_i, \bar{q}).$$

The coverse inequality follows from the inequality

$$\bar{f}(\bar{p},\bar{q}) = \sum_{i=1}^{N}\sum_{j=1}^{M} a_{i,j}p_i q_j = \\ = \sum_{j=1}^{M}\left(\sum_{i=1}^{N} a_{i,j}p_i\right) q_j \geqslant \\ \geqslant \left(\min_j \sum_{i=1}^{N} p_i a_{i,j}\right)\left(\sum_{j=1}^{M} q_j\right) = \\ = \min_j \sum_{i=1}^{N} p_i a_{i,j} = \min_j \bar{f}(\bar{p}, 1_j)$$

that holds for each $\bar{q}$. This inequality means that the minimum of the weighted linear combination is achieved when all of the weight is concentrated on the smallest element. Hence,

$$\min_{\bar{q}} \bar{f}(\bar{p},\bar{q}) \geqslant \min_j \bar{f}(\bar{p}, 1_j).$$

The second inequality is proved similarly. $\triangle$

This theorem implies

Corollary 6.1. *For the mixed extension of an arbitrary matrix game the following equality holds:*

$$v = \max_{\bar{p}} \min_j \bar{f}(\bar{p}, 1_j) = \min_{\bar{q}} \max_i \bar{f}(1_i, \bar{q}),$$

where v is the value of the game.

Let us find a solution of matching pennies game in mixed strategies. The matrix of this game is

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The mixed strategies of this game are $\bar{p} = (p, 1-p)$ and $\bar{q} = (q, 1-q)$,

and the mathematical expectation of the payoff function is

$$\bar{f}(\bar{p},\bar{q}) = \sum_{i,j=1}^{2} a_{i,j}p_iq_j = \\ = q(-p+1-p) + (1-q)(p-(1-p)) = \\ = -4\left(p-\frac{1}{2}\right)\left(q-\frac{1}{2}\right).$$

This expectation

$$\bar{f}(\bar{p},\bar{q}) = -4\left(p-\frac{1}{2}\right)\left(q-\frac{1}{2}\right) \tag{6.10}$$

is the equation of a one-sheet hyperboloid.

Let

$$v(p) = \min_j \bar{f}(\bar{p},1_j) = \min\{1-2p, 2p-1\}.$$

By Corollary 6.1 the value of this game is equal to the maximum of $v(p)$ that is attained for $p^* = \frac{1}{2}$.

Similar arguments show that $q^* = \frac{1}{2}$. the value of the game is $v^* = \bar{f}(\bar{p}^*,\bar{q}^*) = 0$. The point (p^*,q^*) is a saddle point of the one-sheet hyperboloid (6.10).

6.3.3. Solution of the matrix game of type $(2 \times M)$

To find a solution in the mixed extension of a matrix game of type $(2 \times M)$ we will use a geometric representation of its strategies. By Corollary 6.1 the value of this game is equal to

$$v = \max_p \min_{1\leqslant j\leqslant M}(a_{1,j}p + a_{2,j}(1-p)).$$

Here, the first player chooses a mixed strategy – the probability distribution $\bar{p} = (p, 1-p)$ on the rows of the matrix, and the second player chooses a pure strategy – jth column of the matrix.

So, to find the value of the game, the first player must simply find the maximum point $p = p^*$ of the function

$$v(p) = \min_{1\leqslant j\leqslant M}(a_{1,j}p + a_{2,j}(1-p))$$

on the unit interval $[0,1]$.

To find the solution of the game consider all M stright lines

$$L_j(p) = a_{1,j}p + a_{2,j}(1-p),$$

where $j = 1,\ \ldots,\ M$.

For each $p \in [0,1]$ conduct a vertical line to the intersection with the line with the lowest value of the ordinate. The points of intersection form the broken line $y = v(p)$ – *lower envelope* for all these lines. The upper point of the lower envelope determines optimal strategy for the first player (its abscissa is p^*) and the value of the game (ordinate of $v(p^*)$).

Problem. Find the solution of the mixed extension of the matrix game:

$$\begin{pmatrix} 7 & 3 & 3 & 1 & -1 & 0 \\ -1 & -1 & 1 & 0 & 5 & 3 \end{pmatrix}.$$

We build all the lines of the form $L_j(p) = a_{1,j}p + a_{2,j}(1-p)$ for $j = 1,\ \ldots,\ 6$:

$$\begin{aligned} L_1(p) &= 7p - (1-p), \\ L_2(p) &= 3p - (1-p), \\ L_3(p) &= 3p + (1-p), \\ L_4(p) &= p, \\ L_5(p) &= -p + 5(1-p), \\ L_6(p) &= 3(1-p). \end{aligned}$$

We build the lower envelope of these lines. Point p^* is the point of intersection of lines 4 and 5, ie, we solve the equation $p = -p+5(1-p)$. We obtain: $p^* = 5/7$ and $v(p^*) = 5/7$.

To find the optimal strategy of the second player use the following theorem.

Theorem 6.6. *Let $(\bar{p}^*, \bar{q}^*)$ be a solution of a matrix game in mixed strategies, v^* be the value of this game. Then*

- $\bar{f}(1_i, \bar{q}^*) = v^*$ *follows from* $p_i^* > 0$,
- $\bar{f}(\bar{p}^*, 1_j) = v^*$ *follows from* $q_j^* > 0$.

Proof. Let us prove the first statement. By definition $\bar{f}(1_i, \bar{q}^*) \leqslant v^*$, $i = 1, \ \ldots, \ N$.

Assume that a number i_0 exists such that $p^*_{i_0} > 0$ and $\bar{f}(1_{i_0}, \bar{q}^*) < v^*$. Consider the linear combination of inequalities $\bar{f}(1_i, \bar{q}^*) \leqslant v^*$ with coefficients p^*_i, $i = 1, \ \ldots, \ N$, and, since one of the summed inequalities is strict, we get

$$v^* = \bar{f}(\bar{p}^*, \bar{q}^*) = \sum_{i=1}^{N} \bar{f}(1_i, \bar{q}^*) p^*_i < v^* = \bar{f}(\bar{p}^*, \bar{q}^*).$$

This proves the first assertion. $\triangle$

Corollary 6.2. *Let $(\bar{p}^*, \bar{q}^*)$ be a solution of the matrix game in mixed strategies and v^* be a value of the game. Then*

- $p^*_i = 0$ *follows from* $\bar{f}(1_i, \bar{q}^*) < v^*$,
- $q^*_j = 0$ *follows from* $\bar{f}(\bar{p}^*, 1_j) > v^*$.

Condition $\bar{f}(\bar{p}^*, 1_j) = pa_{1,j} + (1-p)a_{2,j} > v^*$ means that the corresponding line at point p^* is above the point of intersection of (two) lines on which the value of the game is attained.

Now complete the solution of the problem - we find the optimal strategy of the second player.

For the 1st, 2nd, 3rd, and 6th pure strategies of the second player (the corresponding lines) we have

$$\bar{f}(\bar{p}^*, 1_j) = L_j(p^*) > v^*$$

for $j = 1, 2, 3, 6$.

By Corollary 6.2, for optimal strategy

$$\bar{q}^* = (q^*_1, q^*_2, q^*_3, q^*_4, q^*_5, q^*_6)$$

we have $q^*_1 = 0$, $q^*_2 = 0$, $q^*_3 = 0$, $q^*_6 = 0$, $q^*_4 = q$, $q^*_5 = 1 - q$.

Now suppose that the first player chooses a pure strategy on the rows – one of the row $i = 1, 2$. The second player chooses a mixed strategy $\bar{q}^* = (0, 0, 0, q^*_4, 1 - q^*_4, 0)$ on the columns. Then

$$v^* = \min_q \max_{1 \leqslant i \leqslant 2} (a_{i,4} q + a_{i,5}(1-q)) = \max_{1 \leqslant i \leqslant 2} (a_{i,4} q^*_4 + a_{i,5}(1 - q^*_4)).$$

For $j = 4, 5$, we have $q_4^* - (1 - q_4^*) = 5/7$ and $5(1 - q_4^*) = 5/7$. Then $q_4^* = 6/7$, $q_5^* = 1/7$.

The complete solution of the game is given by

$$\bar{p}^* = \left(\frac{5}{7}, \frac{2}{7}\right),$$
$$\bar{q}^* = (0, 0, 0, \frac{6}{7}, \frac{1}{7}, 0),$$
$$v^* = \frac{5}{7}.$$

6.3.4. Solution of the game of type $(N \times M)$

Consider a game in mixed strategies with a matrix $A = (a_{i,j})$, where $i = 1, \ \ldots, \ N$, $j = 1, \ \ldots, \ M$. Without loss of generality, suppose that all elements of the matrix A are strictly positive, so the value v of this game are also strictly positive. [1]

By Corollary 6.1 the following equality holds for the mixed extension of matrix game.

$$v = \max_p \min_j \bar{f}(\bar{p}, 1_j) = \min_{\bar{q}} \max_i \bar{f}(1_i, \bar{q}), \tag{6.11}$$

where v is the value of the game. From this follows, that there exists a mixed strategy $\bar{p} = (p_1, \ \ldots, \ p_N)$ of the first player such that $\bar{f}(\bar{p}, 1_j) \geqslant v$ for each pure strategy 1_j of the second player. In other words,

$$\sum_{i=1}^{N} a_{i,j} p_i \geqslant v \quad j = 1, \ \ldots, \ M,$$
$$\sum_{i=1}^{N} p_i = 1,$$
$$p_i \geqslant 0 \quad i = 1, \ \ldots, \ N.$$

[1] In order to achieve this, you can add some sufficiently large positive constant to each element of the payoff matrix.

Denote $x_i = p_i/v$, $i = 1, \ \ldots, \ M$. Then these conditions can be rewritten as

$$\sum_{i=1}^{N} a_{i,j} x_i \geqslant 1 \quad j = 1, \ \ldots, \ M,$$

$$\sum_{i=1}^{N} x_i = 1/v,$$

$$x_i \geqslant 0 \quad i = 1, \ \ldots, \ N.$$

Therefore, the problem of finding solutions of the matrix game is reduced to a linear programming problem: find $x_1, \ \ldots, \ x_N$ such that

$$\sum_{i=1}^{N} x_i \to \min$$

subject to

$$\sum_{i=1}^{N} a_{i,j} x_i \geqslant 1 \text{ for } j = 1, \ \ldots, \ M,$$

$$x_i \geqslant 0 \text{ for } i = 1, \ \ldots, \ N.$$

By (6.11) there is a mixed strategy $\bar{q} = (q_1, \ \ldots, \ q_N)$ of the first player such that $\bar{f}(1_i, \bar{q}) \leqslant v$ for any pure strategy 1_i the first player. In other words, the following conditions are valid:

$$\sum_{j=1}^{M} a_{i,j} q_j \leqslant v \text{ for } i = 1, \ \ldots, \ N,$$

$$\sum_{j=1}^{M} q_j = 1,$$

$$q_j \geqslant 0 \text{ for } j = 1, \ \ldots, \ M.$$

We introduce the notation: $x'_j = q_j/v$, $j = 1, \ \ldots, \ M$. Then these

conditions are transformed into relations

$$\sum_{j=1}^{M} a_{i,j}x'_j \leqslant 1 \text{ for } i = 1,\ \ldots,\ N,$$

$$\sum_{j=1}^{M} x'_j = 1/v,$$

$$x'_j \geqslant 0 \text{ for } j = 1,\ \ldots,\ M.$$

The problem of searching solutions in the matrix game is reduced to a linear programming problem: find $x'_1,\ \ldots,\ x'_M$ such that

$$\sum_{j=1}^{M} x'_j \to \max$$

subject to

$$\sum_{j=1}^{M} a_{i,j}x'_j \leqslant 1 \text{ for } i = 1,\ \ldots,\ N,$$

$$x'_j \geqslant 0 \text{ for } j = 1,\ \ldots,\ M.$$

This is a linear programming problem dual to the direct problem for the variables x_i, $i = 1,\ \ldots,\ N$.

6.3.5. Finite game between K players

In general, a finite game between K players in *normal form* is defined as follows. Player $k \in \{1,\ \ldots,\ K\}$ has N_k possible strategies (moves or pure strategies). Let $\bar{i} = (i_1,\ \ldots,\ i_K)$ be a set of strategies of K players, where $i_j \in \{1,\ \ldots,\ N_j\}$, $j = 1,\ \ldots,\ K$.

Then the gain of the kth player is denoted $f^k(\bar{i}) = f^k(i_1,\ \ldots,\ i_K)$ (in other setting $f^k(\bar{i})$ is his loss).

A mixed strategy of kth player is a probability distribution $\bar{p}^k = (p^k_1,\ \ldots,\ p^k_{N_k})$ in the set of its strategies $\{1,\ \ldots,\ N_k\}$. Here p^k_j is the probability of a pure strategy $j \in \{1,\ \ldots,\ N_k\}$.

Let I^k be a random variable taking any value $i \in \{1,\ \ldots,\ N_k\}$ with probability p^k_i.

Let $\bar{I} = (I^1, \ \ldots, \ I^K)$ be a vector-valued random variable representing strategies of all players. Its values are vectors $\bar{i} = (i_1, \ \ldots, \ i_K)$, where $i_j \in \{1, \ \ldots, \ N_j\}$, $j = 1, \ \ldots, \ K$.

It is usually assumed that the random variables $I^1, \ \ldots, \ I^K$ are independent. Also, a probability measure $\pi = \bar{p}^1 \times \cdots \times \bar{p}^K$ is considered on the set of vectors $\bar{I}$ which determines the probability of elementary event $\bar{i} = (i_1, \ \ldots, \ i_K)$ equal to the product of the probabilities of outcomes:

$$\pi(\bar{i}) = \pi(\bar{I} = \bar{i}) = p^1_{i_1} \cdot \ \ldots \ \cdot p^K_{i_K}.$$

The expected payoff of the kth player is equal to

$$E_\pi(f^k(\bar{I})) = \sum_{\bar{i}} \pi(\bar{i}) f^k(\bar{i}) =$$

$$= \sum_{i_1=1}^{N_1} \cdots \sum_{i_K=1}^{N_K} p^1_{i_1} \cdot \ \ldots \ \cdot p^K_{i_K} f^k(i_1, \ \ldots, \ i_K).$$

Nash equilibrium

The set of mixed strategies of all K players

$$\pi = (\bar{p}^1, \ \ldots, \ \bar{p}^k, \ \ldots, \ \bar{p}^K)$$

is called *Nash equilibrium* if for any $k = 1, \ \ldots, \ K$ and for any mixd strategy $\bar{p}'^k$

$$E_\pi(f^k) \geqslant E_{\pi'}(f^k),$$

where the strategy

$$\pi' = (\bar{p}^1, \ \ldots, \ \bar{p}'^k, \ \ldots, \ \bar{p}^K)$$

is obtained from the strategy π by replacing the probability distribution $\bar{p}^k$ on the probability distribution $\bar{p}'^k$.

We can say that if π is a Nash equilibrium, then there is no advantage to any player to change its strategy if the other players do not change their strategies.

Minimax theorem is a special case of assertion of the existence of the Nash equilibrium for the case of zero-sum games for two players.

In this case, the payoff function of players are $f^1(i,j) = f(i,j)$ and $f^2(i,j) = -f(i,j)$, where $f(i,j)$ is the payoff function of two-person zero-sum game.

In particular, the saddle point $(\bar{p}^0, \bar{q}^0)$ in the two-person zer-sum game in mixed strategies is a Nash equilibrium, since for all mixed strategies $\bar{p}$ and $\bar{q}$

$$\bar{f}(\bar{p}, \bar{q}^0) \leqslant \bar{f}(\bar{p}^0, \bar{q}^0) \leqslant \bar{f}(\bar{p}^0, \bar{q}),$$

where $\bar{f}(\bar{p}, \bar{q})$ is the mathematical expectation of the gain of the first player and $-\bar{f}(\bar{p}, \bar{q})$ is the mathematical expectation of the gain of the second player.

In the case of two-person zero-sum game the set of all Nash equilibria is described in the following proposition.

Proposition 6.1. *A pair if mixed strategies $(\bar{p}^*, \bar{q}^*)$ is a Nash equilibrium in two-person zero-sum game if and only if*

$$\bar{q}^* \in \{\bar{q} : \min_{\bar{p}} \bar{f}(\bar{p}, \bar{q}) \to \max\},$$
$$\bar{p}^* \in \{\bar{p} : \max_{\bar{q}} \bar{f}(\bar{p}, \bar{q}) \to \min\}.$$

For any such pair $(\bar{p}^*, \bar{q}^*)$, the equality $\bar{f}(\bar{p}^*, \bar{q}^*) = v$ holds, where v is a value of the game.

Proof of Proposition 6.1 is left to the reader as a problem.

In general, for finite game of K players the following theorem holds.

Theorem 6.7. *Any finite game has at least one Nash equilibrium.*

The proof of this theorem is based on Brouwer's fixed point theorem.

We give examples of games and Nash equilibria. Consider the games of two players, each of which has two strategies.

Example 1. The first game - previously considered matching pennies game, in which the first player thinks of a number 0 or 1, and the second guesses, with the payoff matrix

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

This zero-sum game has no saddle point, but it has a solution in mixed strategies: the corresponding mixed strategies for the first and second players are $\bar{p}^* = (\frac{1}{2}, \frac{1}{2})$ and $\bar{q}^* = (\frac{1}{2}, \frac{1}{2})$. This solution and is the unique Nash equilibrium in this game.

We rewrite the payoff matrix of this game in more general form:

action	0	1
0	(-1,1)	(1,-1)
1	(1,-1)	(-1,1)

Example 2. The two players decide to go to a concert to listen to Bach or go to a concert to listen to Penderecki. One prefers to listen to Bach, and another to Penderecki. At the same time, they both prefer to go together for one show, than each in his own show. Table of preferences is:

action	Bach	Penderecki
Bach	(2,1)	(0,0)
Penderecki	(0,0)	(1,2)

There are two Nash equilibrium in pure strategies in this game (B,B) and (P,P).

Example 3. Two people live in the neighboring rooms. Everyone can to listen to loud or soft music. Each of them prefers to listen to loud music, and that his neighbor was listening to soft music. Table preferences degree of loudness is:

action	soft	loud
soft	(3,3)	(1,4)
loud	(4,1)	(2,2)

There is only one Nash equilibrium in this game. This is a pure strategy (s, s) (proof is left to a reader as the problem).

Correlated equilibrium

The correlated equilibrium of Aumann is a generalization of the Nash equilibrium. The probability distribution P on the set

$$\prod_{k=1}^{K} \{1, \ \ldots, \ N_k\}$$

of all possible tuples $\bar{i} = (i_1, \ \ldots, \ i_K)$, composed of various strategies of all K players is called *correlated equilibrium* if for all $k = 1, \ \ldots, \ K$ and for any function $h : \{1, \ \ldots, \ N_k\} \to \{1, \ \ldots, \ N_k\}$,

$$E_P(f^k(\bar{i})) \geqslant E_P(f^k(\bar{i}_{-k}, h(i_k))), \tag{6.12}$$

where the vector $\bar{i} = (i_1, \ \ldots, \ i_K)$ is distributed according tothe probability distribution P, and

$$\bar{i}_{-k} = (i_1, \ \ldots, \ i_{k-1}, i_{k+1}, \ \ldots, \ i_K)$$
$$(\bar{i}_{-k}, h(i_k)) = (i_1, \ \ldots, \ i_{k-1}, h(i_k), i_{k+1}, \ \ldots, \ i_K).$$

Unlike the Nash equilibrium values i_k are no longer assumed to be independent, and the probability measure P is not a product measure of mixed strategies of the players.

The following lemma gives an equivalent description of the correlated equilibrium in geometric terms.

Lemma 6.2. *A probability distribution P on the set*

$$\prod_{k=1}^{K} \{1, \ \ldots, \ N_k\}$$

of sequences of strategies $\bar{i} = (i_1, \ \ldots, \ i_K)$ is a correlated equilibrium if and only if for any player $k \in \{1, \ \ldots, \ K\}$ and for any strategies $j, j' \in \{1, \ \ldots, \ N_k\}$

$$\sum_{\bar{i}: i_k = j} P(\bar{i})(f^k(\bar{i}) - f^k(\bar{i}_{-k}, j') \geqslant 0, \tag{6.13}$$

where $(\bar{i}_{-k}, j') = (i_1, \ \ldots, \ i_{k-1}, j', i_k, \ \ldots, \ i_K)$.

The condition (6.13) can be written also as

$$E(f^k(\bar{i})|i_k = j) \geqslant E(f^k(\bar{i}_{-k}, j')|i_k = j), \tag{6.14}$$

where E is the conditional mathematical expectation with respect to the probability distribution P. [2]

[2] Often it is convenient to take this condition as a definition of the correlated equilibrium.

Proof. The condition (6.12) the correlated equilibrium is equivalent to a set of conditions:

$$\sum_{\bar{i}} P(\bar{i})(f^k(\bar{i}) - f^k(\bar{i}_{-k}, h(i_k))) \geqslant 0, \tag{6.15}$$

where $k \in \{1,\ \ldots,\ K\}$ and h is any function of the form

$$h : \{1,\ \ldots,\ N_k\} \to \{1,\ \ldots,\ N_k\}.$$

For any $j, j' \in \{1,\ \ldots,\ N_k\}$, consider a function h such that $h(j) = j'$ and $h(i_k) = i_k$ for all $i_k \neq j$.

Then in the sum (6.15) will be only the terms corresponding to sets $\bar{i}$, where $i_k = j$, and in the remaining terms of the corresponding differences will be reduced. Thus, the sum (6.15) becomes the sum (6.13).

The converse statement is trivial. $\triangle$

Let P be a probability distribution on the set $\prod_{k=1}^{K}\{1,\ \ldots,\ N_k\}$ and $a \in A_k$ for some $1 \leqslant k \leqslant K$. We denote by $P_{-i}(\cdot|i_k = a)$ the corresponding conditional probability distribution on the set $\prod_{s=1,s\neq k}^{K}\{1,\ \ldots,\ N_s\}$ of tuples $\bar{i}_{-k}$ given $i_k = a$. We also introduce the notation

$$f^k(j, \bar{P}_{-k}(\cdot|i_k = a))) = E_{\bar{P}_{-k}(\cdot|i_k=a))}(f^k(j, \bar{i}_{-k}))$$

that is the mathematical expectation of the payoff function, in which $i_k = a$, with respect to this conditional distribution.

We also write more compactly:

$$f^k(j, \bar{P}_{-k}) = E_{\bar{P}_{-k}}(f^k(j, \bar{i}_{-k})),$$

having in mind that $\bar{P}_{-k}$ is a probability distribution on $\bar{i}_{-k}$ generated by the distribution P, provided $i_k = a$.

We can now write the condition (6.13) of the correlated equilibrium in the equivalent form:

Corollary 6.3. *A probability distribution P on the set $\prod_{k=1}^{K}\{1,\ \ldots,\ N_k\}$ of sequences of strategies of type $\bar{i} = (i_1,\ \ldots,\ i_K)$ is correlated equilibrium if and only if for each player $k \in \{1,\ \ldots,\ K\}$ and for each strategy $j, j' \in \{1,\ \ldots,\ N_k\}$*

$$f^k(j, \bar{P}_{-k}(\cdot|i_k = j)) = \max_{j' \in A_i} f^k(j', \bar{P}_{-k}(\cdot|i_k = j)). \tag{6.16}$$

Each condition of the type (6.13) defines a closed half-plane, so that the set of correlated equilibria is a closed convex polyhedron in the space of measures on the set $\prod_{k=1}^{K}\{1,\ \ldots,\ N_k\}$.

The existence of a Nash equilibrium in any finite game means that a correlated equilibrium exists in any finite game. The set of correlated equilibria is a broader and has a simpler description than the set of all Nash equilibria.

6.4. Problems

1. Prove that in the mixed extension of any arbitrary matrix game maximin (minimax) strategy of one player achieved with pure strategy of another player:

$$\begin{aligned} \min_{\bar{q}} \bar{f}(\bar{p}^*, \bar{q}) &= \min_{j} \bar{f}(\bar{p}^*, 1_j), \\ \max_{\bar{p}} \bar{f}(\bar{p}, \bar{q}^*) &= \max_{i} \bar{f}(1_i, \bar{q}^*), \end{aligned} \tag{6.17}$$

where $(\bar{p}^*, \bar{q}^*)$ is solution of the game (saddle point).

2. Prove Proposition 6.1.

3. Prove that in the game of Example 2 (Section 6.3.5), there is also a Nash equilibrium in mixed strategies: the first player chooses B with probability $\frac{2}{3}$ and P with probability $\frac{1}{3}$, and the second player chooses B with probability $\frac{1}{3}$ and P - with probability $\frac{2}{3}$.

Are there any other Nash equilibrium in this game?

4. Prove that in the game of Example 3 (Section 6.3.5), there is only one Nash equilibrium. This is a pure strategy (T, T).

5. Show that an arbitrary convex combination of Nash equilibria is a correlated equilibrium.

Chapter 7

Game-theoretic approach to probability theory

In this chapter we consider a new game-theoretic approach to probability theory, proposed by Vovk and Shafer [26].

Within this approach we formulate games in which, under certain conditions, various laws of probability theory hold. Examples of such laws – the law of large numbers, the law of iterated logarithm, central limit theorem, etc.

Game theory interpretation of probability proposed in Vovk and Shafer's book [26] will be demonstrated in Section 7.1 for the law of large numbers.

Within this approach in the most natural way the problem of universal prediction discussed in Chapter 3 is formulated. Games for universal predictions will be considered in Section 7.3.

7.1. Game-theoretic law of large numbers

Game interpretation of the theory of probability is based on ideas and concepts from finance. In the game-theoretic setting, Vovk and Shafer [26] formulate for every law of probability theory (for example, for the strong law of large numbers and the law of the iterated logarithm) a repeated game with perfect information, in which at each round (step) of the game one participant – *Forecaster*, gives the esti-

mated value of a future outcome and, after that, another participant – *Nature*, issues a new outcome. [1] The third party of the game – *Skeptic* defines the goal of the game. Knowing the forecast, *Skeptic* bets on on its deviation from a future outcome and receives a gain or suffers loss when the outcomes occurs.

Before the game starts *Skeptic* has some initial capital and throughout the game, he can not go to the debt – its strategy should be *defended*. The game is designed so that if the law of probability theory is violated for some sequence of forecasts and outcomes, then, using some specific strategy, *Skeptic* can increase his capital to infinity regardless of the other player moves. This is equivalent to that for the sequence of forecasts and outcomes, for what the law is valid, *Skeptic's* capital will always be bounded.

Consider an infinitely repeated *bounded* game of prediction between three players: *Forecaster*, *Skeptic* and *Nature*.

The players are regulated the following perfect-information protocol:

Initialize the *Skeptic's* capital: $\mathcal{K}_0 = 1$.
FOR $n = 1, 2, \ \ldots$
Forecaster announces a forecast $p_n \in [0, 1]$.
Skeptic announces a number $M_n \in \mathcal{R}$.
announces an outcome $\omega_n \in [0, 1]$.
updates his *capital*: $\mathcal{K}_n = \mathcal{K}_{n-1} + M_n(\omega_n - p_n)$.
ENDFOR

This game can be considered as financial process. In this game, at each step n, *Skeptic* buys M_n units of a financial instrument by p_n per unit. At the end of the step n, $\widehat{\text{N}}$ature announces a new price ω_n and *Skeptic's* capital increases or decreases by the corresponding value. Note that it can be $M_n < 0$. In this case, *Skeptic* sells the number M_n of units of the instrument at the beginning of step n.

Skeptic wins in this game if $\mathcal{K}_n \geqslant 0$ for all n and $\sup \mathcal{K}_n = \infty$ regardless of the other player moves, otherwise *Nature* and *Predictor* win.

[1] In the case of binary outcomes 0 and 1, the average value is equal to the probability of 1.

A trajectory is a sequence of moves of *Forecaster* and *Nature*: $p_1, \omega_1, p_2, \omega_2, \ldots$. We do not assume that there are laws defining the moves of participants. If such a law exists, call it the strategy. Example of *Skeptic's* strategy: at each step n the value M_n can be determined by the sequence of functions from the preceding part of trajectory:

$$M_n = M_n(p_1, \omega_1, p_2, \omega_2, \ldots\ , p_{n-1}, \omega_{n-1}, p_n).$$

Game-theoretic law of large numbers is formulated in the following theorem.

Theorem 7.1. *A defensive strategy for Skeptic exists such that for any trajectory of the game the following holds: if the strong law of large numbers*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) = 0, \tag{7.1}$$

fails then Skeptic wins in the bounded forecasting game; in more detail, Skeptic can choose his moves M_n such that $\mathcal{K}_n \geqslant 0$ for all n and $\limsup_{n\to\infty} \frac{\ln \mathcal{K}_n}{n} > 0$.

Proof. Assume that the strong law of large numbers (7.1) fails. This means that for some $\epsilon > 0$

$$\frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) > 2\epsilon \tag{7.2}$$

holds for infinitely many n or for some $\epsilon > 0$

$$\frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) < -2\epsilon \tag{7.3}$$

holds for infinitely many n.

Consider the first case. Since $|\omega_i - p_i| \leqslant 1$,

$$\epsilon \sum_{i=1}^{n} (\omega_i - p_i) - \epsilon^2 \sum_{i=1}^{n} (\omega_i - p_i)^2 > \epsilon^2 n$$

holds for infinitely many n. Using the inequality $t - t^2 \leqslant \ln(1+t)$, which holds for all $t \geqslant 1/2$, we obtain

$$\sum_{i=1}^{n} \ln(1 + \epsilon(\omega_i - p_i)) > \epsilon^2 n$$

for infinitely many n.

Define the *Skeptic's* strategy:

$$M_n = \epsilon \mathcal{K}_{n-1},$$

for each n, where $\mathcal{K}_{n-1}$ is its current capital.

Then the *Skeptic's* capital at any step n is equal to

$$\mathcal{K}_n = \prod_{i=1}^{n} (1 + \epsilon(\omega_i - p_i)), \tag{7.4}$$

and its logarithm is

$$\ln \mathcal{K}_n = \sum_{i=1}^{n} \ln(1 + \epsilon(\omega_i - p_i)) > \epsilon^2 n,$$

From this, we obtain

$$\limsup_{n \to \infty} \frac{\ln \mathcal{K}_n}{n} > \epsilon^2, \tag{7.5}$$

ie, $\sup \mathcal{K}_n = \infty$.

Note also, that by definition (7.4), $\mathcal{K}_n \geqslant 0$ for all n regardless of the values of ω_i and p_i be announced by *Nature* and *Forecaster* in the process of the game.

Similarly, if (7.3) holds for infinitely many n, we can use *Skeptic's* strategy

$$M_n = -\epsilon \mathcal{K}_{n-1},$$

for all n, where $\mathcal{K}_{n-1}$ is its current capital.

The drawback of this argument is that the *Skeptic* has no information about which of the conditions (7.2) or (7.3) holds for infinitely many n, and for which ϵ they hold.

In order to overcome this difficulty, we complicate the *Skeptic's* strategy so that it takes into account both cases and all possible values of $\epsilon > 0$. We put $\epsilon_k = 2^{-k}$ for $k = 1, 2, \ldots$. We define $\mathcal{K}_0^{1,k} = 1$ and $\mathcal{K}_0^{2,k} = 1$ for all k.

Consider the sequence of strategies and the corresponding auxiliary games: $k = 1, 2, \ldots,$

$$\begin{aligned} M_n^{1,k} &= \epsilon_k \mathcal{K}_{n-1}^{1,k}, \\ M_n^{2,k} &= -\epsilon_k \mathcal{K}_{n-1}^{2,k}, \\ M_n^+ &= \sum_{k=1}^{\infty} 2^{-k} M_n^{1,k}, \\ M_n^- &= \sum_{k=1}^{\infty} 2^{-k} M_n^{2,k}, \\ M_n &= \frac{1}{2}(M_n^+ + M_n^-), \end{aligned}$$

where $n = 1, 2, \ldots$. Combine these auxiliary games and strategies in one game and one strategy M_n with one common gain $\mathcal{K}_n$:

$$\begin{aligned} \mathcal{K}_n^+ &= \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{1,k}, \\ \mathcal{K}_n^- &= \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{2,k}, \\ \mathcal{K}_n &= \frac{1}{2}(\mathcal{K}_n^+ + \mathcal{K}_n^-), \end{aligned}$$

for $n = 1, 2, \ldots$.

These series are convergent, since for any fixed n it hold $\mathcal{K}_n^{1,k} \leqslant 2^n$ for all k. From this and by definition, $|M_n^{2,k}| \leqslant 2^{n-1}$ for all n.

Note that each of capitals satisfies the inequalities $\mathcal{K}_n^{1,k} \geqslant 0$ and $\mathcal{K}_n^{2,k} \geqslant 0$ for all n and k.

If the strong law of large numbers, if condition (7.1) fails then, for some $\epsilon = \epsilon_k$, the condition (7.2) or the condition (7.3) holds for infinitely many n.

By (7.5), where $\mathcal{K}_n = \mathcal{K}_n^{s,k}$,

$$\limsup_{n\to\infty} \frac{\ln \mathcal{K}_n^{s,k}}{n} > 0$$

for $s = 0$ or 1. From this follows that

$$\limsup_{n\to\infty} \frac{\ln \mathcal{K}_n}{n} > 0.$$

Theorem is proved. $\triangle$

Game-theoretic form of the law of large numbers are obtained by conversion and some weakening of Theorem 7.1.

Corollary 7.1. *A defensive strategy of Skeptic exists such that for any trajectory of the bounded forecasting game the following implication is valid:*

$$\sup_n \mathcal{K}_n < \infty \Rightarrow \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) = 0,$$

where $\mathcal{K}_n$ is Skeptic capital at step n.

We say that *Skeptic forces Forecaster* and *Nature* to satisfy the strong law of large numbers.

7.2. Game-theoretic probability

In the Shafer and Vovk's [26] game-theoretic approach to probability theory, the main concept is that of the game of prediction. The notion of probability of an event is a derivative concept and is defined in the game-theoretic terms.

We consider a game of a very general form. At first, we give an informal explanation. In this game we distinguish two players: *Skeptic* and *Nature*. In the financial interpretation, the second player can be called *Market*. *Skeptic* makes moves changing his capital $\mathcal{K}$. At any round n of the game, a move of *Skeptic* is defined by a price to be paid immediately and a payoff that depends on *Nature's* following move. The gambles among which *Skeptic* may choose may depend on

the situation, but we always allow him to combine allowable gambles and to take any fraction or multiple of any available gamble. We also allow him to borrow money freely without paying interest.

All moves of *Nature* form a trajectory of the game. A strategy for *Skeptic* is a plan for how to gamble in each nonterminal round of the game. This strategy can be represented by a function of an initial part of the trajectory available for *Skeptic* at the corresponding round of the game. The *Skeptic's* strategy $\mathcal{M}$ together with his initial capital determine his capital $\mathcal{K}^{\mathcal{M}}(\xi)$ for every trajectory ξ of the game.

The formal definitions are as follows. At each round (step) n of a game *Nature* and *Skeptic* make their moves: *Nature* announces an element w_n, *Skeptic* announces an element p_n. A trajectory of *Nature* is a sequence of its moves over first $n-1$ rounds of the game:

$$\xi^{n-1} = w_1, w_2, \ldots, w_{n-1}.$$

In general case, we assume that w_n is an element of some set $W_{\xi^{n-1}}$ depending from the initial fragment ξ^{n-1} of the trajectory.

Let Ω be a set of all (finite) trajectories of the game.

By a strategy of *Skeptic* we mean any function $p_n = \mathcal{M}(\xi^{n-1})$ from initial fragment of the trajectory. In general, p_n is an element of some set: $p_n \in S_{\xi^{n-1}}$.

Let *a gain function* $\lambda : W_{\xi^{n-1}} \times S_{\xi^{n-1}} \to \mathcal{R}$ be given. At each round of the game the *Skeptic's* capital changes:

$$\mathcal{K}^{\mathcal{M}}(\xi^n) = \mathcal{K}^{\mathcal{M}}(\xi^{n-1}) + \lambda(w_n, p_n),$$

where $\mathcal{K}^{\mathcal{M}}(\xi^0) = \mathcal{K}_0^{\mathcal{M}}$ is its ititial capital.

Any *Skeptic's* strategy $\mathcal{M}$ and a trajectory ξ of *Nature* define the *Skeptc's* capital $\mathcal{K}^{\mathcal{M}}(\xi)$.

We suppose that all strategies of *Skeptic* form a linear space: $\alpha_1 p + \alpha_2 p' \in S_\xi$ for all $p, p' \in S_\xi$ and for all real numbers α_1, α_2, where ξ is an arbitrary initial fragment of the trajectory. We suppose also that the function λ is linear by the second argument:

$$\lambda(w, \alpha_1 p + \alpha_2 p') = \alpha_1 \lambda(w, p) + \alpha_2 \lambda(w, p')$$

for all p, p' and for all real numbers α_1, α_2. In particular, for any two strategies $\mathcal{M}_1$ and $\mathcal{M}_2$ and for any real numbers α_1 and α_2, any

linear combination $\mathcal{M} = \alpha_1\mathcal{M}_1 + \alpha_2\mathcal{M}_2$ is also a strategy and the corresponding *Skeptic* capitals satisfy:

$$\mathcal{K}^{\mathcal{M}}(\xi) = \alpha_1\mathcal{K}^{\mathcal{M}_1}(\xi) + \alpha_2\mathcal{K}^{\mathcal{M}_2}(\xi)$$

for all trajectories ξ of *Nature*.

The example of such game is the simple forecasting game considered in Section 7.1. In the financial interpretation, we can join *Forecaster* and *Nature* in one player *Nature* or *Market*. In the notation of this game p_i is the price of one unit of a financial instrument and ω_i is a payoff for this unit at the end of round i. The *Nature* move at round n is a pair (p_n, ω_n) and the gain function is $\lambda(w, M_n) = M_n(\omega_n - p_n)$.

The set Ω of all trajectories over N rounds consists of all sequences

$$\xi^N = p_1, \omega_1, p_2, \omega_2, \ldots, p_N, \omega_N.$$

The corresponding *Skeptic's* capital is:

$$\mathcal{K}(\xi^N) = \mathcal{K}_0 + \sum_{i=1}^{N} M_n(\omega_n - p_n),$$

where $\mathcal{K}_0$ is some initial capital. This initial capital can be zero.

Let $x = x(\xi)$ be a function of a trajectory $\xi \in \Omega$ of the game. This function will be called *variable*, by analogy with a random variable in the theory of probabilities.

In the financial interpretation, the variable x is an obligation (contract) to pay $x(\xi)$ units of the currency if the game ended at the trajectory ξ.

Buying obligation α for x means that the buyer pays the seller the value of α when the game starts, and the seller must return to the buyer the value $x(\xi)$ at the end of the game for any trajectory ξ.

Consider the question, at what minimum price α the seller can sell (and the purchaser to buy) a variable x.

The buyer can pay the seller a value α smaller than the possible payoff $x(\xi)$ at the end of the game. In this case, the seller must compensate for this difference through the game, that is, use the amount of α as the initial capital for the game with some strategy $\mathcal{M}$ and get in the end of the game the enough capital (or even more)

to pay $x(\xi)$ by the obligation: $\mathcal{K}_0 = 0$ and $\mathcal{K}^{\mathcal{M}}(\xi) + \alpha \geqslant x(\xi)$ for any trajectory ξ of the game.

In what follows, for any variable y, the inequality $\mathcal{K}^{\mathcal{M}} \geqslant y$ means that $\mathcal{K}^{\mathcal{M}}(\xi) \geqslant y(\xi)$ for all trajectories ξ. In this case the seller is *Skeptic*, which should implement obligations x by hedging in the game.

The upper price of a variable x is the lowest price α, at which *Skeptic* can sell the variable x such that he could implement the obligation to pay $x(\xi)$ for any trajectory of the game using some strategy $\mathcal{M}$:

$$\overline{E}x = \inf\{\alpha : \exists\mathcal{M}(\mathcal{K}^{\mathcal{M}} \geqslant x - \alpha)\}$$

Let us now consider the question at which maximum price α the buyer can buy (and the seller can sell) the variable x. The seller receives from the buyer a value α at the beginning of the game and is obliged to pay $x(\xi)$ at the end of the game if the trajectory ξ would realizable. Now buyer is *Skeptic*, which should compensate for the difference $\alpha - x(\xi)$ by hedging.

The lower price of a variable x is the maximal price α, such that *Skeptic's* strategy $\mathcal{N}$ exists satisfying $\mathcal{K}^{\mathcal{N}} \geqslant \alpha - x$:

$$\underline{E}x = \sup\{\alpha : \exists\mathcal{N}(\mathcal{K}^{\mathcal{N}} \geqslant \alpha - x)\}$$

Selling x for α is the same as buying $-x$ for $-\alpha$. Then $\overline{E}x = -\underline{E}(-x)$. Formally, this means that

$$\begin{aligned}\underline{E}(-x) &= \sup\{\alpha : \exists\mathcal{N}(\mathcal{K}^{\mathcal{N}} \geqslant \alpha + x)\} = \\ &= -\inf\{\alpha : \exists\mathcal{M}(\mathcal{K}^{\mathcal{M}} \geqslant x - \alpha)\} = -\overline{E}x\end{aligned}$$

The protocol is called *coherent* if for every *Skeptic's* strategy $\mathcal{M}$ a trajectory ξ exists where he cannot win anything more than its initial capital: $\mathcal{K}^{\mathcal{M}}(\xi) \leqslant \mathcal{K}_0$.

Proposition 7.1. *If the protocol of the game is coherent then* $\underline{E}x \leqslant \overline{E}x$ *and* $\underline{E}a = \overline{E}a = a$*, where* a *is a variable such that* $a(\xi) = a$ *for all* ξ*.*

Proof. If $\underline{E}x > \overline{E}x$ then the constants $\alpha_1 < \alpha_2$ exist such that

$$\overline{E}x < \alpha_1 < \alpha_2 < \underline{E}x.$$

Also, two strategies $\mathcal{M}_1$ and $\mathcal{M}_2$ exist such that $\mathcal{K}^{\mathcal{M}_1} \geqslant x - \alpha_1$ and $\mathcal{K}^{\mathcal{M}_2} \geqslant \alpha_2 - x$. Then for the strategy $\mathcal{M} = \mathcal{M}_1 + \mathcal{M}_2$ and for the zero initial capital $\mathcal{K}_0 = 0$, the following inequality holds

$$\mathcal{K}^{\mathcal{M}} = \mathcal{K}^{\mathcal{M}_1} + \mathcal{K}^{\mathcal{M}_2} ge \alpha_2 - \alpha_1 > 0.$$

This contradicts the assumption of coherence.

The proof of the second statement to the proposition is left to the reader as a problem. $\triangle$

Note that the upper price is determined by the interests of the seller of the variable, and the lower price is determined by the interests of the buyer of the variable.

If the lower and upper prices of a variable x are equal: $\underline{E}x = \overline{E}x$, we call this common value Ex *the price* of the variable x. In this case define $\underline{V}x = \underline{E}(x - Ex)^2$ and $\overline{V}x = \overline{E}(x - Ex)^2$.

An upper and a lower probability of an event $S \subseteq \Omega$ can be defined for any game-theoretic protocol. Consider the indicator function of the event E:

$$1_S(\xi) = \begin{cases} 1 \text{ if } \xi \in S, \\ 0 \text{ otherwise.} \end{cases}$$

This function is also a variable defined on the trajectories of the game. So it has an upper and a lower price.

The upper probability of an event S is defined:

$$\overline{P}(S) = \overline{E}(1_S).$$

By definition $\overline{P}(S) \leqslant 1$ for any event S, and:

$$\overline{P}(S) = \inf\{\alpha : \mathcal{K}_0 = \alpha \text{ and } \exists \mathcal{M} \forall \xi (\mathcal{K}^{\mathcal{M}}(\xi) \geqslant 1 \text{ if } \xi \in S, \\ \mathcal{K}^{\mathcal{M}}(\xi) \geqslant 0 \text{ if } \xi \notin S)\}.$$

The lower probability of an event S is defined

$$\underline{P}(S) = \underline{E}(1_S).$$

By definition:

$$\underline{P}(S) = \sup\{\alpha : \mathcal{K}_0 = 0 \text{ and } \exists \mathcal{N} \forall \xi (\mathcal{K}^{\mathcal{N}}(\xi) \geqslant \alpha - 1 \text{ if } \xi \in S, \\ \mathcal{K}^{\mathcal{N}}(\xi) \geqslant \alpha \text{ if } \xi \notin S)\}.$$

If the protocol is coherent then

$$0 \leqslant \underline{P}(S) \leqslant \overline{P}(S) \leqslant 1$$

and

$$\underline{P}(S) = 1 - \overline{P}(\Omega \setminus S).$$

As an example, consider the game-theoretic version of the Bernoulli theorem.

We consider the following perfect-information protocol – Bernoulli protocol.

The players are: *Skeptic* and *Nature*.

Let $0 < \epsilon \leqslant 1$ and $\alpha > 0$ be the initial *Skeptic's* capital: $\mathcal{K}_0 = \alpha$.
FOR $n = 1, 2, \ \ldots$
Skeptic announces a number $M_n \in \mathcal{R}$.
Nature announces an outcome $x_n \in [-1, 1]$.
Skeptic updates his capital: $\mathcal{K}_n = \mathcal{K}_{n-1} + M_n x_n$.
ENDFOR

Let us denote $S_N = \sum\limits_{n=1}^{N} x_n$. *Skeptic* wins in this game if $\mathcal{K}_n \geqslant 0$ for $n = 1, \ldots, N$ and $\mathcal{K}_n \geqslant 1$ or $\left|\frac{S_N}{N}\right| < \epsilon$.

Note that this protocol is coherent, since, in response to any *Skeptic's* move M_n, *Nature* can produce an outcome:

$$x_n = \begin{cases} 1 \text{ if } M_n < 0, \\ -1 \text{ otherwise.} \end{cases}$$

It holds $\mathcal{K}_n \leqslant \mathcal{K}_0$ for all n.

Theorem 7.2. *Skeptic has a winning strategy for* $N \geqslant \frac{1}{\alpha\epsilon^2}$. *In addition,*

$$\underline{P}\left\{\left|\frac{S_N}{N}\right| > \epsilon\right\} \leqslant \overline{P}\left\{\left|\frac{S_N}{N}\right| > \epsilon\right\} \leqslant \frac{1}{N\epsilon^2}.$$

Proof. Firstly note that

$$\begin{aligned} S_n^2 &= S_{n-1}^2 + 2x_n S_{n-1} + x_n^2 = \\ &= S_{n-1}^2 + 2x_n S_{n-1} + 1, \end{aligned}$$

where $S_0 = 0$.

Assume that *Skeptic* choose his move using the rule:

$$M_n = \frac{2\alpha S_{n-1}}{N}$$

at each round n of the game. Then

$$\mathcal{K}_N - \mathcal{K}_0 = \frac{\alpha}{N}\sum_{n=1}^{N} 2S_{n-1}x_n =$$
$$= \frac{\alpha}{N}(S_N^2 - N) = \alpha\left(\frac{S_N^2}{N} - 1\right).$$

Hence, $\mathcal{K}_N = \frac{\alpha S_N^2}{N}$. Then

$$\left|\frac{S_N}{N}\right| \leqslant \sqrt{\frac{\mathcal{K}_N}{\alpha N}}. \tag{7.6}$$

By definition, *Skeptic* wins if $\mathcal{K}_N \geqslant 1$. If $\mathcal{K}_N < 1$ and $N > \frac{1}{\alpha\epsilon^2}$ then by (7.6)

$$\left|\frac{S_N}{N}\right| < \epsilon.$$

Therefore, *Skeptic* wins again.

The sum S_N depends on a trajectory of the game: $S_N = S_N(\xi)$. Let us estimate the upper probability of the event that the average value of this sum deviates from zero by more than a number ϵ:

$$E = \left\{\xi : \left|\frac{S_N(\xi)}{N}\right| > \epsilon\right\},$$

consisting of all the trajectories of the game for which the above inequality holds. Its upper probability is:

$$\overline{P}(E) = \inf\{\alpha : \mathcal{K}_0 = \alpha \text{ and } \exists\mathcal{M}\forall\xi(\mathcal{K}_N^{\mathcal{M}}(\xi) \geqslant 1 \text{ if } \left|\frac{S_N(\xi)}{N}\right| > \epsilon,$$
$$\mathcal{K}_N^{\mathcal{M}} \geqslant 0 \text{ if } \left|\frac{S_N}{N}\right| \leqslant \epsilon)\}.$$

As we have just proved, such a strategy $\mathcal{M}$ exists for $\alpha = \frac{1}{N\epsilon^2}$.

Hence,

$$\underline{P}\left\{\left|\frac{S_N}{N}\right| > \epsilon\right\} \leqslant \overline{P}\left\{\left|\frac{S_N}{N}\right| > \epsilon\right\} \leqslant \frac{1}{N\epsilon^2}.$$

Theorem is proved. $\triangle$

7.3. Game of universal forecasting

In this section, we show that in some modification of the game defined in Section 7.1, *Skeptic*, using a defensive strategy, can force *Forecaster* to issue forecasts that are well-calibrated at arbitrary infinite sequence of outcomes issued by *Nature*.

Consider some infinitely repeated deterministic game between three players: *Forecaster*, *Skeptic* and *Nature*.

The players actions are regulated by the following protocol:

Initialize $\mathcal{K}_0 = 1$.
FOR $n = 1, 2, \ \ldots$
Skeptic announces a function $S_n : [0, 1] \to \mathcal{R}$.
Forecaster announces a forecast $p_n \in [0, 1]$.
Nature announces an outcome $\omega_n \in \{0, 1\}$.
Skeptic updates his capital: $\mathcal{K}_n = \mathcal{K}_{n-1} + S_n(p_n)(\omega_n - p_n)$.
ENDFOR

Winners in the infinite deterministic game:

Forecaster wins if *Skeptic's* capital $\mathcal{K}_n$ is bounded at all round of the game; otherwise, *Skeptic* and *Nature* win.

Theorem 7.3. *Skeptic and Nature have winning strategies in the deterministic forecasting game.*

Proof. Indeed, *Skeptic* can define

$$S_n(p) = \begin{cases} 1 \text{ if } p < 0.5, \\ -1 \text{ otherwise.} \end{cases}$$

Nature can define

$$\omega_n = \begin{cases} 1, \ \ p_n < 0.5, \\ 0 \end{cases}$$

at each round n of the game.

In this game, for each round $n > 0$, if $\omega_n = 0$ then $p_n \geqslant \frac{1}{2}$ and, thus, $\omega_n - p_n \leqslant -\frac{1}{2}$ and $S_n(p_n) = -1$; if $\omega_n = 1$ then $p_n < \frac{1}{2}$ and, thus, $\omega_n - p_n \geqslant \frac{1}{2}$ and $S_n(p_n) = 1$. From this follows

$$\mathcal{K}_n \geqslant \mathcal{K}_{n-1} + \frac{1}{2}$$

for all n, and *Skeptic's* capital is unbounded. $\triangle$

In this game, "adversarial" *Nature* uses the *Forecaster's* prediction for defining her outcome.

It turns out that in a randomized version of this game *Forecaster* wins. In the randomized version of the game *Nature* does not know the precise forecast, she knows only the probability distribution according to which this forecast is generated.

Consider an infinitely repeated game between four players: *Forecaster*, *Skeptic*, *Nature* and *Random numbers generator*. Let $\{0,1\}$ be the set of outcomes, $\mathcal{P}\{0,1\}$ be the set of all probability measures on $\{0,1\}$. [2]

The game is regulated by the following perfect-information protocol.

Initialize $\mathcal{K}_0 = 1$ and $\mathcal{F}_0 = 1$.
FOR $n = 1, 2, \ldots$
Skeptic announces a function $S_n : [0,1] \to \mathcal{R}$.
Forecaster announces a probability distribution on the set of all forecasts: $P_n \in \mathcal{P}[0,1]$.
Nature announces an outcome $\omega_n \in \{0,1\}$.
Forecaster announces a test of randomness $f_n : [0,1] \to \mathcal{R}$, which is correct with respect to the measure P_n, ie, such that $\int f_n(p) P_n(dp) \leqslant 0$.
Random numbers generator announces a number $p_n \in [0,1]$.
Skeptic updates his capital: $\mathcal{K}_n = \mathcal{K}_{n-1} + S_n(p_n)(\omega_n - p_n)$.
Forecaster updates his capital: $\mathcal{F}_n = \mathcal{F}_{n-1} + f_n(p_n)$.
ENDFOR

[2] Each measure $Q \in \mathcal{P}\{0,1\}$ is defined by two numbers $(q, 1-q)$, where $q = Q\{1\}$ is the probability of 1.

Protocol defines that information is available for the players in the process of the game. Each player, when choosing its strategy, can use all information that appeared before his move – outcomes, forecasts and strategies.

Restrictions for Skeptic: Skeptic have to choose S_n such that his capital satisfies $\mathcal{K}_n \geqslant 0$ for all n regardless of the moves of all other players.

Restrictions for Forecaster: Forecaster have to choose his moves P_n and f_n such that his capital satisfies $\mathcal{F}_n \geqslant 0$ for all n regardless of the moves of all other players. [3]

Winners in randomized forecasting game:

We assume that the strategies of the players are such that these constraints are satisfied. If a player at least once violate the constraint, then it can not be a winner in the game.

Forecaster wins in this game if (i) his capital $\mathcal{F}_n$ is unbounded or if (ii) the *Skeptic's* capital $\mathcal{K}_n$ is bounded; in all other cases, *Skeptic* and *Nature* win.

The next theorem shows that *Forecaster* has a winning strategy.

Theorem 7.4. *Forecaster has a winning strategy in the randomized forecasting game.*

Proof. At each step n of our game, consider an auxiliary zero-sum game with players *Nature* and *Forecaster* defined as follows.

Forecaster chooses a number $p_n \in [0,1]$ and *Nature* chooses a number $\omega_n \in \{0,1\}$. The *Forecaster's* loss (the *Nature* gain) is equal to

$$F(\omega_n, p_n) = S(p_n)(\omega_n - p_n).$$

[3] Capital $\mathcal{F}_n$ corresponds to the concept of bounded from below supermartingale from probability theory, and $f_n(p)$ corresponds to the supermartingale-difference $\mathcal{F}_n - \mathcal{F}_{n-1}$. Rules of the game require that the $\mathcal{F}_0 = 1$ and $\mathcal{F}_n \geqslant 0$ for all n in the process of the game. Condition $\int f_n(p) P_n(dp) \leqslant 0$ for all n implies $\int \mathcal{F}_n P_n(dp) \leqslant \mathcal{F}_{n-1}$ for all n.

These properties define the concept of supermartingale in probability theory. In our case, these properties should be carried out only for the trajectory of the game.

For any mixed strategy of *Nature* $Q_n \in \mathcal{P}\{0,1\}$, *Forecaster* presents a pure strategy $p_n = Q\{1\}$. [4]

Then mathematical expectation of *Nature's* gain with respect to the mixed strategy Q and the pure strategy p_n is equal to

$$\begin{aligned} F(Q_n, P_n) = Q\{0\}F(0, p_n) + Q\{1\}F(1, p_n) = \\ = Q\{0\}S(p_n)(-p_n) + Q\{1\}S(p_n)(1 - p_n) = \\ = (1 - Q\{1\})S(p_n)(-Q\{1\}) + Q\{1\}S(p_n)(1 - Q\{1\}) = 0. \end{aligned}$$

Thus, $\forall\ Q\ \exists\ P\ F(Q,P) \leqslant 0$ or

$$\sup_Q \inf_P F(Q,P) \leqslant 0. \tag{7.7}$$

In order to apply the minimax theorem, it is necessary to transform this game into a matrix game.

Consider an approximation to the auxiliary game, in which the set of columns corresponding to *Forecaster* moves is finite. For any $\Delta > 0$, choose a finite ϵ-network N_ϵ in the set $[0,1]$ consisting of rational points, such that each point of $[0,1]$ is located at a distance no more than ϵ of one of the points of this set, and such that the lower value of the game does not exceed $\Delta/2$, when *Forecaster* chooses $p_n \in N_\epsilon$.

Such ϵ-net can be chosen, since $|S_n(p)| \leqslant \mathcal{K}_{n-1} \leqslant 2^{n-1}$ is bounded for all p [5] The the inequality (7.7) will be transformed into the inequality

$$\sup_Q \inf_P F(Q,P) \leqslant \Delta/2.$$

By the minimax theorem,

$$\inf_P \sup_Q F(Q,P) = \sup_Q \inf_P F(Q,P) \leqslant \Delta/2.$$

[4] This pure strategy p_n corresponds to the mixed strategy $P_n(p_n) = 1$ and $P_n(r) = 0$ for $r \in [0,1] \setminus \{p_n\}$.

[5] *Skeptic* has to choose $S_n(p)$ such that $\mathcal{K}_n \geqslant 0$ for all n regardless of actions of other players.

Therefore, *Forecaster* has a mixed strategy $P \in \mathcal{P}[0,1]$ concentrated in the set N_ϵ, such that

$$\sup_Q F(Q,P) \leqslant \Delta.$$

This implies that

$$\int S_n(p)(\omega_n - p)P(dp) \leqslant \Delta \tag{7.8}$$

for both values $\omega_n = 0$ and $\omega_n = 1$.

Let E_Δ be a subset of the set $\mathcal{P}[0,1]$ consisting of probability distributions P satisfying the condition (7.8) for $\omega_n = 0$ and $\omega_n = 1$ simultaneously.

The set of measures $\mathcal{P}[0,1]$ can be supplied by the topology of weak convergence. It is well known in the measure theory, that the space $\mathcal{P}[0,1]$ is compact in this topology. Besides, E_Δ is closed in this topology.

Choose a monotonically decreasing to 0 sequence of rational numbers Δ_i, $i = 1, 2, \ldots$. The intersection of an infinite sequence of closed nested subsets of a compact set is non-empty. Hence, $\cap E_{\Delta_i} \neq \emptyset$.

hen a probability measure $P_n \in \cap E_{\Delta_i} \subseteq \mathcal{P}[0,1]$ exists such what

$$\int S_n(p)(\omega_n - p)P_n(dp) \leqslant 0 \tag{7.9}$$

for $\omega_n = 0$ and $\omega_n = 1$.

We now return to our main game. Strategy of *Forecaster* will be to choose at step n the probability distribution P_n, which has been defined in the auxiliary game. *Forecaster's* second move is to choose the test f_n:

$$f_n(p) = S_n(p)(\omega_n - p).$$

Then $\mathcal{F}_n = \mathcal{K}_n$ for all n.

The mean value of the test f_n by the measure P_n does not exceed 0 by (7.9), ie, the test f_n is correct with respect to the measure P_n.

By $\mathcal{F}_n = \mathcal{K}_n$, there will always be one of two things: the *Skeptic's* gain is bounded or the *Forecaster's* gain is unbounded. In both cases, *Forecaster* wins. $\triangle$

We say that *Random number generator* produces random numbers *perfectly* if $\sup_n \mathcal{F}_n < \infty$.

7.4. Randomized well-calibrated forecasting

In this section, we show that *Skeptic*, choosing in a special way their moves $S_n(p)$, can force *Forecaster* to choose their forecasts so that they were well-calibrated for any sequence of outcomes how would *Nature* choose them.

We first consider a simple case, where *Forecaster* presents its predictions such that the some game-theoretic version of the strong law of large numbers holds. The idea of the construction is the same as in Section 7.1.

Let ϵ be an arbitrary positive real number suc that $0 < \epsilon < 1$. Put $\mathcal{K}_0^1 = 1$. Get

$$S_n^1(p) = \epsilon \mathcal{K}_{n-1}^1$$

in the randomized forecasting game defined in Section 7.1. This *Skeptic's* strategy does not depend on *Forecaster's* predictions but depends on *Skeptic's* gain received on steps $< n$.

In this case *Skeptic's* gain at step n is equal to

$$\mathcal{K}_n^1 = \prod_{i=1}^{n} (1 + \epsilon(\omega_i - p_i)), \tag{7.10}$$

where $\omega_1, \ \ldots, \ \omega_n$ is a sequence of outcomes announced by *Nature*, and $p_1, \ \ldots, \ p_n$ is a sequence of forecasts announced by *Forecaster* on steps $1, \ \ldots, \ n$.

Since $|\omega_i - p_i| \leqslant 1$ for all i, $\mathcal{K}_n^1 \geqslant 0$ for all n regardless of the actions of other players, ie, the basic requirement for *Skeptic's* strategy is fulfilled.

By Theorem 7.4 *Forecaster* has a winning strategy in the randomized forecasting game. This means that if *Random number generator* announces random numbers perfectly, ie, $\sup_n \mathcal{F}_n < \infty$, then, regardless of how *Nature* announces her outcomes $\omega_1, \ \ldots, \ \omega_n$, *Forecaster* can produce the forecasts $p_1, \ \ldots, \ p_n$ such that the *Skeptic's* gain $\mathcal{K}_n^1$

is bounded by some number $C > 0$:

$$\prod_{i=1}^{n}(1+\epsilon(\omega_i - p_i)) \leqslant C$$

n. This inequality can be rewritten in the form

$$\begin{aligned}
\sum_{i=1}^{n}\ln(1+\epsilon(\omega_i - p_i)) &\leqslant \ln C,\\
\epsilon\sum_{i=1}^{n}(\omega_i - p_i) - \epsilon^2\sum_{i=1}^{n}(\omega_i - p_i)^2 &\leqslant \ln C,\\
\epsilon\sum_{i=1}^{n}(\omega_i - p_i) &\leqslant \ln C + \epsilon^2 n,\\
\frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) &\leqslant \frac{\ln C}{\epsilon n} + \epsilon \qquad (7.11)
\end{aligned}$$

for all n. Here we have used the inequality $\ln(1+t) \geqslant t - t^2$ for $|t| \leqslant 0.5$.

From this we obtain

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) \leqslant \epsilon. \qquad (7.12)$$

Similarly, getting $\mathcal{K}_0^2 = 1$ and choosing the strategy

$$S_n^2(p) = -\epsilon\mathcal{K}_{n-1}^2,$$

Skeptic can force *Forecaster* to output his predictions such that the following inequality will hold:

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) \geqslant -\epsilon. \qquad (7.13)$$

Both of these strategies can be joined into a single strategy, which provides simultaneous execution of both inequalities (7.12) and (7.13). In this case, the strategies $S_n^1(p)$ and $S_n^2(p)$, and the

corresponding capitals $\mathcal{K}_n^1(p)$, $\mathcal{K}_n^2(p)$ can be considered by *Skeptic* as auxiliary in his calculations.

Skeptic chooses the strategy

$$S_n(p) = \frac{1}{2}(S_n^1(p) + S_n^2(p)).$$

Skeptic's gain at step n is equal to

$$\mathcal{K}_n = \frac{1}{2}(\mathcal{K}_n^1 + \mathcal{K}_n^2).$$

Note that each gain satisfies $\mathcal{K}_n^1 \geqslant 0$ and $\mathcal{K}_n^2 \geqslant 0$ for all n. At first step $S_1(p) = 0$, since $S_1^1(p) = -S_1^2(p)$, then $S_n^1(p)$ and $S_n^2(p)$ diverge, as they determined on the basis their winnings $\mathcal{K}_n^1(p)$ and $\mathcal{K}_n^2(p)$.

Assume that *Random number generator* announces random numbers perfectly, ie, $\sup_n \mathcal{F}_n < \infty$.

Since the cumulative gain $\mathcal{K}_n$ is bounded, both cumulative gains $\mathcal{K}_n^1$ and $\mathcal{K}_n^2$ are also bounded. As it was proven above these implies the inequalities (7.12) and (7.13).

The next step is to construct *Skeptic's* strategy, which provides simultaneous validity of inequalities (7.12) and (7.13) for all $\epsilon > 0$.

To do this, we introduce a sequence $\epsilon_k = 2^{-k}$ for all k. Define $\mathcal{K}_0^{1,k} = 1$ and $\mathcal{K}_0^{2,k} = 1$ for all k. Consider the sequence of strategies:

$$\begin{aligned}
S_n^{1,k}(p) &= \epsilon_k \mathcal{K}_{n-1}^{1,k}, \\
S_n^{2,k}(p) &= -\epsilon_k \mathcal{K}_{n-1}^{2,k}, \\
S_n^{+}(p) &= \sum_{k=1}^{\infty} 2^{-k} S_n^{1,k}(p), \\
S_n^{-}(p) &= \sum_{k=1}^{\infty} 2^{-k} S_n^{2,k}(p), \\
S_n(p) &= \frac{1}{2}(S_n^{+}(p) + S_n^{-}(p)).
\end{aligned}$$

The corresponding gains satisfy conditions:

$$\mathcal{K}_n^+ = \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{1,k},$$
$$\mathcal{K}_n^- = \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{2,k},$$
$$\mathcal{K}_n = \frac{1}{2}(\mathcal{K}_n^+ + \mathcal{K}_n^-).$$

These series are convergent, since for any fixed n, by (7.10), the inequality $\mathcal{K}_n^{1,k} \leqslant 2^n$ holds for all k. Thus, $|S_n^{2,k}(p)| \leqslant 2^{n-1}$ for all n.

Note that each of the gains satisfies $\mathcal{K}_n^{1,k} \geqslant 0$ and $\mathcal{K}_n^{2,k} \geqslant 0$ for all n and k.

Therefore, the uniform boundedness of total gain $\mathcal{K}_n$ implies that the gains $\mathcal{K}_n^{1,k}$ and $\mathcal{K}_n^{2,k}$ are bounded.

As was shown above, the limitations of each of these gains implies the simultaneous fulfillment of limit inequalities (7.12) and (7.13) for all ϵ_k, $k = 1, 2, \ \ldots$

From this, we obtain that the mixed *Skeptic's* strategy forces *Forecaster* to choose the winning strategy – the randomized forecasts – defined by Theorem 7.4 such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) = 0. \tag{7.14}$$

The definition (7.14) of calibration has an obvious drawback. For example, the sequence of forecasts $p_1, p_2, \ \ldots = \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ is well-calibrated for the sequence $\omega_1, \omega_2, \ \ldots = 0, 1, 0, 1, 0, 1, 0, 1, \ \ldots$ of outcomes.

However, if select only members of the sequence of outcomes that have even (or odd) indices, such forecasts will not be well-calibrated for the corresponding subsequence. Therefore necessary to consider additional checking rules for the selection of subsequences.

Let *Nature* announces a sequence of outcomes $\omega_1, \ \omega_2 \ldots$ and *Forecaster* announces predictions $p_1, p_2, \ \ldots$. *Checking rule* is a binary fuction

$$F(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{n-1}, \omega_{n-1}, p_n),$$

defined on sequences of type

$$p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{n-1}, \omega_{n-1}, p_n,$$

where p_n is a *Forecaster* prediction at step n, $n = 1, 2, \ \ldots$, taking values: 0 1.

A sequence of forecasts $p_1, p_2, \ \ldots$ is called *well-calibrated* for a sequence of outcomes $\omega_1, \omega_2, \ \ldots$ with respect to a checking rule $F(p_1, \omega_1, p_2, \omega_2, \ \ldots \ , p_{n-1}, \omega_{n-1}, p_n)$ if

$$\sup_n \sum_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p_i) < \infty$$

or

$$\lim_{n\to\infty} \frac{\sum\limits_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p_i)(\omega_i - p_i)}{\sum\limits_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p_i)} = 0. \quad (7.15)$$

Note that *Nature* announces her outcome ω_n using the history

$$p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{n-1}, \omega_{n-1}, p_n.$$

The main result of the theory of universal prediction claims:

Theorem 7.5. *For every countable sequence F_k, $k = 1, 2, \ \ldots$ of checking rules, a Forecaster's strategy: an algorithm computing predictions P_n given past outcomes and forecasts*

$$p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{n-1}, \omega_{n-1}$$

, exists such that, for any sequence of outcomes $\omega_1, \omega_2, \ \ldots$ announced by Nature, the sequence of predictions $p_1, p_2, \ \ldots$ issued by Random number generator perfectly is well-calibrated for this sequence of outcomes with respect to any checking rule F_k.

Proof. The proof is the next step of complication of the construction defined above. In the construction of strategies $S_n^{1,k}(p)$ and $S_n^{1,k}(p)$, replace the number ϵ_k on $\epsilon_k F_s$, where $k, s = 1, 2, \ \ldots$

Consider an infinite sequence of auxilliary strategies of *Skeptic*:

$$S_n^{1,k,s}(p) = \epsilon_k F_s(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{n-1}, \omega_{n-1}, p)\mathcal{K}_{n-1}^{1,k,s},$$
$$S_n^{2,k,s}(p) = -\epsilon_k F_s(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{n-1}, \omega_{n-1}, p)\mathcal{K}_{n-1}^{2,k,s}.$$

Consider some effective one to one enumeration of all pairs of positive integer numbers (k, s). Let for ith such pair $p(i) =$ and $q(i) = s$. Such enumeration and the corresponding functions $p(i)$ and $q(i)$ can be defined in many different ways. We omit details of such enumeration.

Define

$$S_n^+(p) = \sum_{j=1}^{\infty} 2^{-j} S_n^{1,p(j),q(j)}(p),$$
$$S_n^-(p) = \sum_{j=1}^{\infty} 2^{-j} S_n^{2,p(j),q(j)}(p),$$
$$S_n(p) = \frac{1}{2}(S_n^+(p) + S_n^-(p)).$$

The rest part of the proof is similar to the case where *Skeptic's* strategies were mixed with weights ϵ_k.

Note that the summation in the modified version of (7.11) should only be performed by those i, for which

$$F(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p_i) = 1.$$

In the modified version of (7.11) and in (7.14), to obtain (7.15), we should replace n in the denominator by

$$\sum_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p_i).$$

7.5. Problems

1. Prove that in Theorem 7.1 the condition $\limsup_{n\to\infty} \mathcal{K}_n = \infty$ can be replaced by condition $\lim_{n\to\infty} \mathcal{K}_n = \infty$ (*Hint*: To prove this, instead

of a single strategy $M_n = \epsilon\mathcal{K}_{n-1}$, consider an infinite number of of strategies of the form

$$M_n^C = \begin{cases} M_n \text{ if } \mathcal{K}_{n-1} \leqslant 2^C, \\ 0 \text{ otherwise,} \end{cases}$$

where C is an arbitrary positive integer number.

Thereafter, we consider a mixture of these strategies

$$\tilde{M}_n = \sum_{C=1}^{\infty} 2^{-C} M_n^C.$$

Denote the corresponding *Skeptic's* capital $\mathcal{K}_n^C$.

We must show that for an arbitrary step n the capital of *Skeptic*, which adheres to the strategy $\tilde{M}_n$, is

$$\tilde{\mathcal{K}}_n = \sum_{C=1}^{\infty} 2^{-C} \mathcal{K}_n^C.$$

From this it is easy to see that $\limsup\limits_{n\to\infty} \mathcal{K}_n = \infty$ if and only if $\lim\limits_{n\to\infty} \tilde{\mathcal{K}}_n = \infty$).

2. Prove the following inequalities for the upper prices of any variables x, x_1, x_2 in a game with the coherent protocol:

a) $\overline{E}x \leqslant \sup\limits_{\xi\in\Omega} x(\xi)$.

b) $\underline{E}a = \overline{E}a = a$, where a is a constant.

c) $\overline{E}(x_1 + x_2) \leqslant \overline{E}x_1 + \overline{E}x_2$.

d) $\overline{E}(x + \alpha) = \overline{E}x + \alpha$, where α is a constant.

e) $\overline{E}(\alpha x) = \alpha\overline{E}x$ for $\alpha > 0$.

f) if $x_1 \leqslant x_2$ then $\overline{E}x_1 \leqslant \overline{E}x_2$.

3. State and prove the similar inequalities for lower price $\underline{E}x$ of any variable x.

4. Prove that $0 \leqslant \overline{P}(S) \leqslant 1$ and $\leqslant \underline{P}(S) \leqslant 1$ in any forecasting game.

5. Prove the following inequalities for the upper probability of any events E, E_1 and E_2 in a game with the coherent protocol:

a) $0 \leqslant \underline{P}(E) \leqslant \overline{P}(E) \leqslant 1$.

b) $\underline{P}(\Omega) = \overline{P}(\Omega) = 1$.

c) $\overline{P}(E) = 1 - \underline{P}(\Omega \setminus E)$.

d) $\overline{P}(E_1 \cup E_2) \leqslant \overline{P}(E_1) + \overline{P}(E_2)$.

e) $\underline{P}(E_1 \cap E_2) \geqslant \underline{P}(E_1) + \underline{P}(E_2) - 1$.

f) if $E_1 \subseteq E_2$ then $\overline{P}(E_1) \leqslant \overline{P}(E_2)$.

6. Consider a game from Section 7.2 (Bernoulli protocol) with the outcomes from the set $\{-1, 1\}$. Let this game is performed over N rounds.

a) Prove that the upper and lower probabilities of any fixed trajectory ξ of length N is equal to 2^{-N} and the upper and lower probabilities of any finite set S is equal to $2^{-N}|S|$.

(*Hint.* Let $\mathcal{K}_0 = 0$. To obtain an upper bound for the upper probability of the trajectory ξ, consider the *Skeptic's* strategy: $M_1 = 2^{-N}$ and $M_t = \mathcal{K}_{t-1}$ for $t \geqslant 2$ along the given trajectory, define $M_t = 0$ after the trajectory of the game diverges from the desired trajectory ξ. In this case *Skeptic's* capital will double at each step, until the trajectory of the game coincides with the given trajectory. Capital becomes zero, as soon as the trajectory of the game diverged with a given trajectory.

To obtain a lower bound for the lower probability, get $\alpha = 2^{-N}$ and define the *Skeptic's* strategy: $M_1 = -\alpha$ and $M_t = \mathcal{K}_{t-1}$ for $t \geqslant 2$ along the given trajectory ξ. Also, $\mathcal{K}_0 = 0$. define $M_t = 0$ after the trajectory of the game diverged from the desired trajectory ξ.

While the trajectory of the game coincides with the given trajectory ξ *Skeptic's* debt at the end of step s is equal to $-\sum_{i=0}^{s-1} 2^i\alpha = -(2^s - 1)\alpha$. If the trajectory of the game diverges from ξ at step s then *Skeptic* wins $\alpha 2^s$ and after paying the debt his capital is α. If the trajectory of the game coincides with ξ the *Skeptic's* debt at the end of the game is $-(2^N - 1)\alpha$, ie, the *Skeptic's* capital at the end of the game is is $\alpha - 1$).

b) Provide examples of events for which you can calculate exactly the upper and lower probabilities.

7. Let in the simple forecasting game from Section 7.2 the outcomes x_i are in the set $\{1, 2\}$. Let also the game is performed over N rounds.

a) Provide examples of events for which you can calculate exactly

these probabilities. Consider $S = \{1^N\}$, $S = \{2^N\}$, $S = \{1^N, 2^N\}$, where k^N – is a sequence consisting of N numbers k.

8. Complete the proof of Theorems 7.1 and 7.5.

Chapter 8

Infinitely repeated games

In Chapter 6 we have considered onetime realizations of different games and evaluate their performance. The calculation of the equilibrium points in these games is computationally time consuming procedures. In particular, we have to solve a linear programming problem to find the points of equilibrium.

In this chapter, using the theory of the well-calibrated forecasting, we show that it is possible to approximate the points of the Nash equilibrium or points of the correlated equilibrium in infinitely repeated games by means of the frequency distributions of players' moves.

In Section 8.1, we consider the asymptotic characteristics of infinitely repeated zero-sum game, and show that the previously constructed machine learning algorithms approximate points of the Nash equilibrium of these games.

In Section 8.2, we prove the Blackwell approachability theorem, which is a generalization of the minimax theorem for the case of vector-valued payoff functions.

In Section 8.3, we apply this theorem to construct well-calibrated predictions for the case of arbitrary finite number of outcomes.

In Section 8.4 we show that if, in some infinitely repeated game, each player uses predictions that are well-calibrated for the sequence of strategies chosen by his opponents, and chooses "the optimal response" for these predictions, then the joint frequency distribution of

the players' strategy converges to the set of correlated equilibria of this game.

8.1. Infinitely repeated two players zero-sum games

In this section we consider game repeatable over time.

Assume that at each step $t = 1, 2, \ldots$ the first player chooses a move $I_t \in \{1, \ldots, N\}$ according to the probability distribution $\bar{p}_t = (p_{1,t}, \ldots, p_{N,t})$ (mixed strategy) and the second player chooses a move $J_t \in \{1, \ldots, M\}$ according to the probability distribution of $\bar{q}_t = (q_{1,t}, \ldots, q_{N,t})$. The mixed strategies of the players $\bar{p}_t$ and $\bar{q}_t$ may depend on the preceding moves of these players and their results.

The gain of the first player at step t is $\bar{f}(\bar{p}_t, \bar{q}_t)$ and the gain of the second player is $-\bar{f}(\bar{p}_t, \bar{q}_t)$.

We will compare the cumulative gain of each player over n steps with the cumulative gain of its best constant strategy:

$$\max_{i=1,\ldots,N} \sum_{t=1}^{n} f(i, J_t) - \sum_{t=1}^{n} f(I_t, J_t)$$

for the first player and

$$\sum_{t=1}^{n} f(I_t, J_t) - \min_{j=1,\ldots,M} \sum_{t=1}^{n} f(I_t, j)$$

for the second player.

We apply the theory predictions with expert advice to approximate the equilibrium in such games.

When analyzing the actions of the first player, the set of his strategies $\{1, \ldots, N\}$ will be considered as a set of auxiliary experts. Each expert i produces the constant prediction $i \in \{1, \ldots, N\}$ at all steps.

The first player is considered as *Forecaster*, which announces at each step t a prediction I_t. Any strategy $J_t \in \{1, \ldots, M\}$ of the second expert is interpreted as an outcome announced by *Nature* at step t.

Similarly, when analyzing the actions of the second player, the set of his strategies $\{1, \ldots, M\}$ will be considered as a set of auxiliary experts. Each expert j produces the constant prediction $j \in \{1, \ldots, M\}$ at all steps.

The second player is considered as *Forecaster*, which announces at each step t a prediction J_t. Any move $I_t \in \{1, \ldots, N\}$ of the first player is interpreted as an outcome announced by *Nature* at step t.

Now explain what the loss function used in this analysis. The loss of the first player is equal $\lambda^1(J_t, I_t) = -f(I_t, J_t)$, where J_t is an outcome announced by *Nature*, and I_t is a forecast announced by the first player at step t. The loss of the second player is equal $\lambda^2(I_t, J_t) = f(I_t, J_t)$, where I_t is an outcome announced by *Nature*, and J_t is a forecast announced by the second player at step t.

The first (or second) player can choose his moves (mixed strategies) according to some rule or algorithm that at each step t outputs a probability distribution $\bar{p}_t$ (or $\bar{q}_t$). Any algorithm of this kind will be called *online strategy* of the first (or second) player in *the infinitely repeated game*.

Assume that both players choose their moves according to Hannan consistent online strategies (see (4.60)). For example, we can use the exponentially weighted forecaster defined in Sections 4.6.

According to this algorithm, at any step t, the first player chooses his mixed strategy $\bar{p}_t = (p_{1,t}, \ldots, p_{N,t})$ by the rule:

$$p_{i,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \lambda^1(i, J_s)\right)}{\sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} \lambda^1(k, J_s)\right)}, \tag{8.1}$$

where $i = 1, \ldots, N$, η_t is a variable learning rate.

At the same time, the strategy J_s of the second player is considered as an outcome announced by *Nature*.

By Corollary 4.3 the first player is Hannan consistent, ie, with probability one,

$$\limsup_{n\to\infty} \left(\frac{1}{n} \sum_{t=1}^{n} \lambda^1(J_t, I_t) - \min_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} \lambda^1(J_t, i) \right) \leqslant 0 \tag{8.2}$$

for a suitable choice of the parameters η_t, where the trajectory $I_1, I_2, \ldots$ is distributed according to the probability distribution generated by the sequence of mixed strategies (8.1).

Note that (8.2) holds regardless of the second player moves. We suppose here that the second player is oblivious, ie, the trajectory $J_1, J_2, \ldots$ is given in advance and does not depend on the first player moves.

The second player can also can apply a similar online strategy. In this case he also is Hannan consistent, ie, with probability one,

$$\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}\lambda^2(I_t,J_t)-\min_{j=1,\ldots,\,M}\frac{1}{n}\sum_{t=1}^{n}\lambda^2(I_t,j)\right)\leqslant 0, \qquad (8.3)$$

where the trajectory $J_1, J_2, \ldots$ is distributed according to the probability distribution generated by the sequence of mixed strategies similar to (8.1). Here we use the similar assumptions on the first player.

In terms of payoff functions (8.2) has a form: with probability one,

$$\liminf_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}f(I_t,J_t)-\max_{i=1,\ldots,\,N}\frac{1}{n}\sum_{t=1}^{n}f(i,J_t)\right)\geqslant 0, \qquad (8.4)$$

where the trajectory $I_1, I_2, \ldots$ is distributed by the measure $\bar{p}_1\times\bar{p}_2\times\ldots$ that is the product of the first player's mixed strategies.

The inequality (8.3) can be rewritten: with probability one:

$$\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}f(I_t,J_t)-\min_{j=1,\ldots,\,M}\frac{1}{n}\sum_{t=1}^{n}f(I_t,j)\right)\leqslant 0, \qquad (8.5)$$

where the trajectory $J_1, J_2, \ldots$ is distributed by the measure $\bar{q}_1\times\bar{q}_2\times\ldots$ that is the product of the second player's mixed strategies.

The following theorem asserts that if the first player chooses his move according to a Hannan consistent online strategy then, regardless on what strategy the second player uses, the average gain of the first player can not be much less than the value the game.

A similar assertion holds for the second player – if the second player chooses his move according to a Hannan consistent online strategy then, regardless on what the first player chooses his moves, the average gain of the second player can not be much more than the value the game.

Theorem 8.1. *Assume that in a two-person zero-sum game the first player chooses his moves according to a Hannan consistent online strategy. Then*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} f(I_i, J_t) \geqslant v, \tag{8.6}$$

alsost surely, regardless of the second player's moves, where v is the value of the game.

If each player uses a Hannan consistent online strategy then, with probability 1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} f(I_i, J_t) = v, \tag{8.7}$$

where the sequence $I_1, J_1, I_2, J_2, \ldots$ is distributed according to probability distribution $\bar{p}_1 \times \bar{q}_1 \times \bar{p}_2 \times \bar{q}_2 \times \ldots$.

Proof. By the minimax theorem the value of the game is equal

$$v = \max_{\bar{p}} \min_{\bar{q}} \bar{f}(\bar{p}, \bar{q}) = \min_{\bar{q}} \max_{\bar{p}} \bar{f}(\bar{p}, \bar{q}).$$

Also, define

$$\bar{f}(\bar{p}, \bar{q}) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_i q_j f(i, j),$$

$$\bar{f}(\bar{p}, j) = \sum_{i=1}^{N} p_i f(i, j),$$

$$\bar{f}(i, \bar{q}) = \sum_{j=1}^{M} q_j f(i, j).$$

By (8.4) to prove the first statement (8.6) it is sufficient to show that for any sequence $J_1, J_2, \ldots,$

$$\max_{i=1,\ldots,\ N} \frac{1}{n} \sum_{t=1}^{n} f(i, J_t) \geqslant v \tag{8.8}$$

for all n. For the proof note that

$$\max_{i=1,\ldots,\ N} \frac{1}{n} \sum_{t=1}^{n} f(i, J_t) = \max_{\bar{p}} \frac{1}{n} \sum_{t=1}^{n} \bar{f}(\bar{p}, J_t),$$

since $\sum\limits_{t=1}^{n} \bar{f}(\bar{p}, J_t)$ is linear form by $\bar{p}$, and the maximum of a linear function defined on the simplex of probability distributions on $\{1, \ldots,\ N\}$ is attained at some its vertex.

Let

$$\hat{q}_{j,n} = \frac{1}{n} \sum_{t=1}^{n} 1_{\{J_t = j\}}$$

be a frequency of rounds where the second player chooses the strategy j. Let also $\hat{q}_n = (\hat{q}_{1,n}, \ldots,\ q_{M,n})$. Then

$$\max_{\bar{p}} \frac{1}{n} \sum_{t=1}^{n} \bar{f}(\bar{p}, J_t) = \max_{\bar{p}} \sum_{j=1}^{M} \hat{q}_{j,n} \bar{f}(\bar{p}, j) =$$
$$= \max_{\bar{p}} \bar{f}(\bar{p}, \hat{q}_n) \geqslant \min_{\bar{q}} \max_{\bar{p}} \bar{f}(\bar{p}, \bar{q}) = v$$

for any sequence $J_1, J_2, \ldots.$

To prove the second assertion (8.7) of the theorem we use condition (8.5) of the Hannan consistency, and show that

$$\min_{j=1,\ldots,\ M} \frac{1}{n} \sum_{t=1}^{n} f(I_t, j) \leqslant v = \max_{\bar{p}} \min_{\bar{q}} \bar{f}(\bar{p}, \bar{q})$$

for any sequence $I_1, I_2, \ldots.$ This proof is similar to the proof of the inequality (8.8).

From this we obtain

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} f(I_i, J_t) \leqslant v, \tag{8.9}$$

almost surely, where v is the value of the game.

Combining (8.9) and (8.6), we obtain (8.7). Theorem is proved. $\triangle$

8.2. Blackwell approachability theorem

Theorem 8.1 of the previous section states that the first player using a Hannan consistent online strategy at a sufficiently large number of steps, can do the mean value of his gain asymptotically less than the value of the game, no matter what strategy the second player uses.

In this section we consider a generalization of this result for the case of a vector-valued payoff function and an arbitrary closed convex set S instead of the value of the game. We will prove the famous Blackwell approachability theorem. In 1956, Blackwell [6] proposed a generalization of the minimax theorem for the case of a vector-payoff function. Later it was observed that this theorem can be used to construct the well-calibrated forecasts.

This theorem provides the necessary and sufficient conditions under which there exists a randomized online strategy of the first player such that, with probability 1, for an unlimited continuation of the game, he can approximate the mean value of the payoff vector to a given set S, regardless of the second player moves.

As well as before we consider a two persons game. Only now the payoff function $f(i, j)$ takes values in the d-dimensional space $\mathcal{R}^d$.

Recall that the strategies of the first player belong to a finite set $\mathcal{I} = \{1, \ldots, N\}$, and the strategies of the second player belong to a finite set $\mathcal{J} = \{1, \ldots, M\}$. Mixed strategies of the players are probability distribution in the sets $\mathcal{I}$ and $\mathcal{J}$. The sets of these mixed strategies are denoted $\mathcal{P}(\mathcal{I})$ and $\mathcal{P}(\mathcal{J})$, correspondingly. Recall

notations:

$$f(\bar{p},j)=\sum_{i=1}^{N}p_i f(i,j),$$
$$f(i,\bar{q})=\sum_{j=1}^{M}q_j f(i,j),$$
$$f(\bar{p},\bar{q})=\sum_{i=1}^{N}\sum_{j=1}^{M}p_i q_j f(i,j)$$

for $\bar{p}=(p_1,\dots,p_N)\in\mathcal{P}(\mathcal{I})$ and $\bar{q}=(q_1,\dots,q_M)\in\mathcal{P}(\mathcal{J})$.

We consider the Euclidean distance

$$\|\bar{x}-\bar{y}\|=\sqrt{\sum_{i=1}^{d}(x_i-y_i)^2}$$

between two vectors $\bar{x},\bar{y}\in\mathcal{R}^d$. For any subset $S\subseteq\mathcal{R}^d$ and any vector $\bar{x}\in\mathcal{R}^d$, the distance from $\bar{x}$ to S is defined

$$\operatorname{dist}(\bar{x},S)=\inf_{\bar{y}\in S}\|\bar{x}-\bar{y}\|.$$

For any closed set S, let $d_S(\bar{x})$ denotes an element $\bar{y}\in S$ such that the distance $\operatorname{dist}(\bar{x},\bar{y})$ is minimal. If the set S is convex, then this element is unique.

A set $S\subseteq\mathcal{R}^d$ is called *approachable* if a randomized online strategy $\bar{p}_1,\bar{p}_2,\dots$ of the first player exists such that for any sequence $J_1,J_2,\dots$ of the second player moves

$$\lim_{T\to\infty}\operatorname{dist}\left(\frac{1}{T}\sum_{t=1}^{T}f(I_t,J_t),S\right)=0$$

holds for P-almost all sequences $I_1,I_2,\dots$ of the first player moves, where $P=\prod\bar{p}_t$ is the overall probability distribution on trajectories $I_1,I_2,\dots$ of the first player moves generated by its mixed strategies $\bar{p}_1,\bar{p}_2,\dots.$

The following theorem gives a sufficient condition for the approachability of a closed convex subset of $\mathcal{R}^d$.

Assume that the set S and the values of $f(i,j)$ are located in the unit ball of the space $\mathcal{R}^d$.

Theorem 8.2. *Let a closed subset $S \subseteq \mathcal{R}^d$ be given. For each vector $\bar{x} \notin S$, consider the hyperplane $\Pi_{\bar{x}}$ passing through $d_S(\bar{x})$ and orthogonal to the line passing through $\bar{x}$ and $d_S(\bar{x})$.*

Assume that for every vector $\bar{x} \notin S$ there is a probability distribution $\bar{p} \in \mathcal{P}(\mathcal{I})$ such that the points $f(\bar{p},1),\ldots,\ f(\bar{p},M)$ and the point $\bar{x}$ lie on different sides of the hyperplane $\Pi_{\bar{x}}$.

Then the set S is approachable.

Proof. Let $I_1, I_2, \ldots$ and $J_1, J_2, \ldots$ be some strategies of the first and the second players. Let

$$\bar{m}_t = \frac{1}{t}\sum_{i=1}^{t} f(I_i, J_i)$$

be the average gain of the first player over first t steps.

Assume that $\bar{m}_t \notin S$ and, at steps $< t$ of the game, the players performed the moves $I_1,\ldots,\ I_{t-1}$ and $J_1,\ldots,\ J_{t-1}$. Equation of the hyperplane $\Pi_{\bar{x}}$ passing through the point $d_S(\bar{m}_{t-1})$ and orthogonal to the line connecting points $\bar{m}_{t-1}$ and $d_S(\bar{m}_{t-1})$ has the form:

$$(\bar{w}_{t-1} \cdot \bar{x}) - b_{t-1} = 0,$$

where

$$\bar{w}_{t-1} = \frac{\bar{m}_{t-1} - d_S(\bar{m}_{t-1})}{\|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|}$$

and

$$b_{t-1} = (\bar{w}_{t-1} \cdot d_S(\bar{m}_{t-1})).$$

Suppose that $\bar{m}_0 = \bar{0}$.

Note that the point $\bar{m}_{t-1}$ is above of the hyperplane (since it is the end of the direction vector of the hyperplane).

By assumption of the theorem for the point $\bar{x} = \bar{m}_{t-1}$ a mixed strategy $\bar{p}_t$ of the first player exists such that all the points

$$f(\bar{p}_t,1),\ldots,\ f(\bar{p}_t,M)$$

are below of this hyperplane:

$$(\bar{w}_{t-1} \cdot f(\bar{p}_t, j)) - b_{t-1} \leqslant 0$$

for all $j = 1, \ldots, M$. We rewrite this condition in the form:

$$\max_{1 \leqslant j \leqslant M} (\bar{w}_{t-1} \cdot (f(\bar{p}_t, j) - d_S(\bar{m}_{t-1}))) \leqslant 0. \tag{8.10}$$

A mixed strategy $\bar{p}_t$ is a solution of the linear programming problem (8.10).

We verify that the point $\bar{m}_t$ is "approaching" to the set S. By definition

$$\operatorname{dist}(\bar{m}_t, S) = \|\bar{m}_t - d_S(\bar{m}_t)\| \leqslant \|\bar{m}_t - d_S(\bar{m}_{t-1})\|. \tag{8.11}$$

It is easy to verify that

$$\bar{m}_t = \frac{t-1}{t}\bar{m}_{t-1} + \frac{1}{t} f(I_t, J_t). \tag{8.12}$$

Square the inequality (8.11), and perform the calculations using the equality (8.12):

$$\begin{aligned}
\operatorname{dist}(\bar{m}_t, S)^2 &\leqslant \left\| \frac{t-1}{t}\bar{m}_{t-1} + \frac{1}{t} f(I_t, J_t) - d_S(\bar{m}_{t-1}) \right\|^2 = \\
= \left\| \frac{t-1}{t}(\bar{m}_{t-1} - d_S(\bar{m}_{t-1})) + \frac{1}{t}(f(I_t, J_t) - d_S(\bar{m}_{t-1})) \right\|^2 &= \\
= \left(\frac{t-1}{t} \right)^2 \|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|^2 + & \\
+2\frac{t-1}{t^2}((\bar{m}_{t-1} - d_S(\bar{m}_{t-1})) \cdot (f(I_t, J_t) - d_S(\bar{m}_{t-1}))) + & \\
+\frac{1}{t^2}\|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|^2. &
\end{aligned} \tag{8.13}$$

Since the set S and all values $f(i, j)$ are located in the unit ball of the space $\mathcal{R}^d$, the following inequality holds:

$$\|f(I_t, J_t) - d_S(\bar{m}_{t-1})\| \leqslant 2.$$

Using this inequality, transform the inequalities (8.12) and (8.13) into the inequality

$$t^2\|\bar{m}_t - d_S(\bar{m}_t)\|^2 - (t-1)^2\|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|^2 \leqslant$$
$$\leqslant 4 + 2(t-1)((\bar{m}_{t-1} - d_S(\bar{m}_{t-1})) \cdot (f(I_t, J_t) - d_S(\bar{m}_{t-1}))). \quad (8.14)$$

Denote

$$K_{t-1} = \frac{t-1}{T}\|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|.$$

We have $0 \leqslant K_{t-1} \leqslant 2$ for $t \leqslant T$. Summing the left and the right parts of the inequality (8.14) over $t = 1, \dots, T$ and dividing it by T^2, we obtain:

$$\|\bar{m}_T - d_S(\bar{m}_T)\|^2 \leqslant$$
$$\leqslant \frac{4}{T} + \frac{2}{T}\sum_{t=1}^{T} K_{t-1}(\bar{w}_{t-1} \cdot (f(I_t, J_t) - d_S(\bar{m}_{t-1}))) \leqslant$$
$$\leqslant \frac{4}{T} + \frac{2}{T}\sum_{t=1}^{T} K_{t-1}(\bar{w}_{t-1} \cdot (f(I_t, J_t) - f(\bar{p}_t, J_t))). \quad (8.15)$$

To obtain the last inequality we have used the inequality (8.11).

The second term of the last member of (8.15) is a martingale-difference. [1] Therefore, by Corollary 8.7 (Azuma–Hoeffding inequality), it tends to 0 almost surely as $T \to \infty$. Then

$$\operatorname{dist}(\bar{m}_T, S) = \|\bar{m}_T - d_S(\bar{m}_T)\| \to 0 \text{ as } T \to \infty$$

with probability 1. Theorem is proved. $\triangle$

The following theorem gives the necessary and sufficient conditions under which an arbitrary closed convex set is approachable by the first player.

Theorem 8.3. *A closed convex subset $S \subseteq \mathcal{R}^d$ is approachable by the first player if and only if for every mixed strategy $\bar{q} \in \mathcal{P}(\mathcal{J})$ a mixed strategy $\bar{p} \in \mathcal{P}(\mathcal{I})$ exists such that $f(\bar{p}, \bar{q}) \in S$.*

[1] Indeed, $E_{\bar{p}_t}(f(I_t, J_t)) = f(\bar{p}_t, J_t)$, where E is the symbol of mathematical expectation.

Proof. Assume that for every $\bar{q} \in \mathcal{P}(\mathcal{J})$ an $\bar{p} \in \mathcal{P}(\mathcal{I})$ exists such that $f(\bar{p}, \bar{q}) \in S$. Let also $\bar{x}_0 \notin S$ and $d_S(\bar{x}_0)$ be the point of S closest to the point $\bar{x}_0$.

Consider the auxiliary matrix game with the payoff function $a(i,j) = ((d_S(\bar{x}_0) - \bar{x}_0) \cdot f(i,j))$. By the minimax theorem

$$\max_{\bar{p}} \min_{j} a(\bar{p}, j) = \min_{\bar{q}} \max_{i} a(i, \bar{q}). \tag{8.16}$$

By assumption of the theorem 8.3 for every $\bar{q} \in \mathcal{P}(\mathcal{J})$ a number i exists such that $f(i, \bar{q}) \in S$. From this and by (8.16) we obtain

$$\begin{aligned} &\max_{\bar{p}} \min_{j}((d_S(\bar{x}_0) - \bar{x}_0) \cdot f(\bar{p}, j)) = \\ &= \min_{\bar{q}} \max_{\bar{p}}((\bar{d}_S(\bar{x}_0) - \bar{x}_0) \cdot f(i, \bar{q})) \geqslant \\ &\geqslant \min_{\bar{s} \in S}((d_S(\bar{x}_0) - \bar{x}_0) \cdot \bar{s}) = \\ &= ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(\bar{x}_0)). \end{aligned} \tag{8.17}$$

The last inequality of (8.17) follows from the definition of ${}_S(\bar{x}_0)$.

Consider the hyperplane

$$L(\bar{x}) = ((d_S(\bar{x}_0) - \bar{x}_0) \cdot \bar{x}) - ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(x_0)) = 0$$

passing through $d_S(\bar{x}_0)$ and orthogonal to the vector $d_S(\bar{x}_0) - \bar{x}_0$. It is easy to verify that

$$((d_S(\bar{x}_0) - \bar{x}_0) \cdot x_0) < ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(\bar{x}_0)).$$

Then $L(\bar{x}_0) < 0$, ie, he point $\bar{x}_0$ is below the hyperplane $L(\bar{x}) = 0$.

By the inequality between the first and the last terms of (8.17) a mixed strategy $\bar{p} \in \mathcal{P}(\mathcal{I})$ exists such that for all $j = 1, \ldots, M$:

$$((d_S(\bar{x}_0) - \bar{x}_0) \cdot f(\bar{p}, j)) \geqslant ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(\bar{x}_0)).$$

In other words, $L(f(\bar{p}, j)) \geqslant 0$ for all $j = 1, \ldots, M$, ie, the hyperplane $L(\bar{x}) = 0$ separates these points and the point $\bar{x}_0$. Hence, the set S is approachable by Theorem 8.2.

To prove the converse statement, assume that an $\bar{q}_0 \in \mathcal{P}(\mathcal{J})$ exists such that $f(\bar{p}, \bar{q}_0) \notin S$ for all $\bar{p} \in \mathcal{P}(\mathcal{I})$.

We apply Theorem 8.2 for the game with the transposed payoff matrix (payoff function) $f'(i,j) = f(j,i)$ and a closed convex set $T(\bar{q}_0) = \{f(\bar{p}, \bar{q}_0) : \bar{p} \in \mathcal{P}(\mathcal{I})\}$.

By definition $f'(\bar{q}_0, 0), \ldots, f'(\bar{q}_0, N) \in T(\bar{q}_0)$. By convexity of the set $T(\bar{q}_0)$ for every $\bar{x} \notin T(\bar{q}_0)$ the points $\bar{x}$ and $f'(\bar{q}_0, 0), \ldots, f'(\bar{q}_0, N)$ are on the opposite sides of the hyperplane $\Pi_{\bar{x}}$. Then by Theorem 8.2 the set $T(\bar{q}_0)$ is approachable for the second player using a constant strategy $\bar{q}_0$ and the transposed matrix $f'(i,j)$.

We have supposed that $T(\bar{q}_0) \cap S = \emptyset$. By the assumption the sets S and $T(q_0)$ are closed. It is easy to see that the set S cannot be aproachable by the first player (see a problem in Section 8.5). Theorem is proved. $\triangle$

As a first application of Theorem 8.2 we construct a Hannan consistent forecasting (online) strategy.

Let $\mathcal{I} = \{1, \ldots, N\}$ be a set of all strategies of the first player and $\mathcal{J} = \{1, \ldots, M\}$ be a set of all strategies moves of the second player, $\mathcal{P}(\mathcal{I})$ and $\mathcal{P}(\mathcal{J})$ be sets of their mixed strategies.

Note that it is not important in the Blackwell approachability theorem that type of function: payoff or loss, is used. In this application we consider a loss function $l(i,j)$, where $0 \leqslant l(i,j) \leqslant 1$ for all i, j.

Our goal is to define at each step t a mixed strategy $\bar{p}_t$ of the first player such that for any sequence of moves $J_1, J_2, \ldots$ of the second player

$$\limsup_{T \to \infty} \left(\frac{1}{T} \sum_{t=1}^{T} l(I_t, J_t) - \min_{1 \leqslant i \leqslant N} \frac{1}{T} \sum_{t=1}^{T} l(i, J_t) \right) \leqslant 0, \qquad (8.18)$$

with probability 1, where moves $I_1, I_2, \ldots$ of the first player are distributed according to the product distribution $\prod_t \bar{p}_t$.

In order to apply Theorem 8.2, we consider a closed convex set

$$S = \{(u_1, \ldots, u_N) : u_i \leqslant 0, i = 1, \ldots, N\},$$

and a vector-valued payoff function

$$\bar{f}(i,j) = \begin{pmatrix} l(i,j) - l(1,j) \\ \dots \\ l(i,j) - l(k,j) \\ \dots \\ l(i,j) - l(N,j) \end{pmatrix}.$$

The values of $f(i,j)$ are located in an N-dimensional ball of radius $\sqrt{N}$ centered in the origin. Multiplying this function by a constant $1/\sqrt{N}$ we can ensure that the values of $f(i,j)$ lie in the unit ball.

Let $\bar{x}_0 \notin S$. Is sufficient to consider the case where $d_S(x_0) = \bar{0}$ and the equation of the hyperplane $\Pi_{\bar{x}_0}$ has the form $(\bar{w} \cdot \bar{x}) = 0$, where all the components of w_i of the normal vector $\bar{w}$ of the hyperplane are nonnegative.

To prove the existence of a strategy such that (8.18) holds, it suffices to prove that there is a mixed strategy $\bar{p} \in \mathcal{P}(\mathcal{I})$ such that all vectors $f(\bar{p},1), \dots, f(\bar{p},M)$ lie below the hyperplane $(\bar{w} \cdot \bar{x}) = 0$, ie, the following inequality

$$\sum_{k=1}^{N} w_k (l(\bar{p},j) - l(k,j)) \leqslant 0$$

holds for all $j = 1, \dots M$. It is easy to verify that this condition holds for

$$\bar{p} = \frac{\bar{w}}{\sum\limits_{i=1}^{N} w_i}.$$

By Theorem 8.2 a sequence of mixed strategies $\bar{p}_1, \dots, \bar{p}_t, \dots$ exists such that condition (8.18) holds with probability 1.

8.3. Calibrated forecasting

In this section, we present a method for constructing calibrated forecasts on the basis of Theorem 8.3 for the case of arbitrary finite set of outcomes. This method was proposed by Mannor and Stoltz [24].

In Section 3.2 we have considered the problem of universal prediction of the mean value p_i of a future outcome ω_i and the corresponding concept of calibration. In this section we consider a problem of universal prediction of the probability distribution of future outcomes. In the case of binary set of outcomes $\{0, 1\}$ these two problems are equivalent, since the probability p_i of $\omega_n = 1$ is the mean value of a future outcome $\omega_i \in \{0, 1\}$.

We assume that the outcomes are elements of a finite set $A = \{a_1, \ldots, a_m\}$. Denote by $\mathcal{P}(A)$ the set of all probability distributions in the set A. Any such distribution (mixed strategy) is a vector $\bar{p} = (p_1, \ldots, p_m)$, where the sum of all its coordinates is equal 1. We consider a norm $\|\bar{p}\|_1 = \max_{1 \leqslant i \leqslant m} |p_i|$ on the set of all such vectors. The Euclidian norm $\|\bar{p}\|_2$ in $\mathcal{R}^m$ is also suitable. It is known that these norms are equivalent in $\mathcal{R}^m$. In what follows $\|\bar{p}\|$ denotes any such norm.

Let $\bar{\delta}[a_i] = (0, \ldots, 1, \ldots, 0)$ be a probability distribution concentrated on element a_i of the set A. In this vector, the ith coordinate is 1, all other coordinates are 0.

We consider a perfect information game between two players: *Forecaster* and *Nature*. At each step t *Forecaster* announces a probability distribution $\bar{p}_t \in \mathcal{P}(A)$, after that, *Nature* announces an outcome $a_t \in A$.

In terms of the game theory, $\bar{p}_t$ is a mixed strategy of *Forecaster* and $\bar{\delta}[a_t]$ is a pure strategy of *Nature*.

We also consider probability distributions in the set of all mixed strategies that are probability distributions in the set of probability distributions $\mathcal{P}(A)$. This set is denoted $\mathcal{P}(\mathcal{P}(A))$.

For selecting strategies $\bar{p}_1, \bar{p}_2, \ldots$, *Forecaster* will use randomization, more precisely, *Forecaster* will issue at each step t a random vector $\bar{p}_t \in \mathcal{P}(A)$ distributed according to some probability distribution $\bar{P}_t \in \mathcal{P}(\mathcal{P}(A))$.

By the Ionescu-Tulcea [29] theorem the probability measures P_t, $t = 1, 2, \ldots$, can be regarded as conditional distributions with respect to a overall distribution $P = \prod P_t$ defined on trajectories $\bar{p}_1, \bar{p}_2, \ldots$.

Each player can use all the information known to each his action. There are no restrictions for the strategy of *Nature*.

Let a real number $\epsilon > 0$ be given. *Forecaster's* goal is to output randomized forecasts $\bar{p}_t$ ditributed by the measure P such that for any $\bar{p} \in \mathcal{P}(A)$ and for any sequence of *Nature* moves $a_1, a_2, \ldots$, P-almost surely the condition of ϵ-calibration holds:

$$\limsup_{T\to\infty} \left\| \frac{1}{T} \sum_{t=1}^{T} I_{\|\bar{p}_t - \bar{p}\| \leqslant \epsilon} (\bar{p}_t - \bar{\delta}[a_t]) \right\| \leqslant \epsilon, \tag{8.19}$$

where vectors $\bar{p}_1, \bar{p}_2, \ldots$ are distributed by the measure P and

$$I_{\|\bar{p}_t - \bar{p}\| \leqslant \epsilon} = \begin{cases} 1 \text{ if } \|\bar{p}_t - \bar{p}\| \leqslant \epsilon, \\ 0 \text{ otherwise.} \end{cases}$$

Forecaster will choose the forecasts $\bar{p}_t$ from a fixed finite set of mixed strategies

$$\mathcal{P}_\epsilon = \{\bar{s}_1, \ldots, \bar{s}_N\} \subset \mathcal{P}(A).$$

To specify this set we construct some ϵ-net in the set $\mathcal{P}(A)$ of all mixed strategies, which are m-dimensional vectors. Thus, for any vector $\bar{p} \in \mathcal{P}(A)$ an element $\bar{s}_i \in \mathcal{P}_\epsilon$ exists such that $\|\bar{p} - \bar{s}_i\| < \epsilon$.

We will define the probability distributions $P_t \in \mathcal{P}(\mathcal{P}(A))$ concentrated in a finite set $\mathcal{P}_\epsilon$ of mixed strategies.

For simplicity, we identify the finite set $\mathcal{P}_\epsilon = \{\bar{s}_1, \ldots, \bar{s}_N\}$ and the set of indexes of its elements $\mathcal{I} = \{1, 2, \ldots, N\}$. We will also consider at each step t probability distributions P_t in $\mathcal{I}$.

The overall probability distribution on trajectories $i_1, i_2, \ldots$ of these indices is defined $P = \prod P_t$. Then the condition (8.19) follows from the condition: P-almost surely,

$$\limsup_{T\to\infty} \sum_{k=1}^{N} \left\| \frac{1}{T} \sum_{t=1}^{T} I_{\{i_t = k\}} (\bar{s}_k - \bar{\delta}[a_t]) \right\| \leqslant \epsilon, \tag{8.20}$$

where the trajectories $i_1, i_2, \ldots$ are distributed by the measure P.

The existence of ϵ-calibrated strategy in general form is asserted in the following theorem.

Theorem 8.4. *For any $\epsilon > 0$, a probability distribution P can be constructed such that P-almost surely the condition of ϵ-calibration*

(8.19) holds for each $\bar{p} \in \mathcal{P}(A)$, where the vectors $\bar{p}_1, \bar{p}_2, \ldots$ are distributed by P. [2]

Proof. We apply Theorem 8.3, in which the first player is considered as *Forecaster* using strategies from a set [3] $\mathcal{I} = \{1, 2, \ldots, N\}$, and the second player is considered as *Nature* using the set of strategies $\mathcal{J} = A$. The values of payoff function are vectors of dimension $N|A|$:

$$f(k,a) = \begin{pmatrix} \bar{0} \\ \ldots \\ \bar{0} \\ \bar{s}_k - \bar{\delta}[a] \\ \bar{0} \\ \ldots \\ \bar{0} \end{pmatrix}.$$

where $k \in \mathcal{I}$ and $a \in \mathcal{J}$, $\bar{0}$ is the m-dimensional zero vector, $m = |A|$, and $\bar{s}_k - \bar{\delta}[a]$ are difference of two m-dimensional column vectors, which is kth component of the complex vector $f(k,a)$.

We now define a convex set in the space $\mathcal{R}^{mN}$. We consider vectors in $\mathcal{R}^{mN}$ as complex vectors of dimension N with the vector components from $\mathcal{R}^m$: $\bar{X} = (\bar{x}_1, \ldots, \bar{x}_N)$, where $\bar{x}_i \in \mathcal{R}^m$.

We define the closed convex set of such complex vectors:

$$C = \left\{ \bar{X} : \sum_{k=1}^{N} \|\bar{x}_k\| \leqslant \epsilon \right\}.$$

By Theorem 8.3 the closed convex set C is approachable if and only if for each $\bar{q} \in \mathcal{P}(\mathcal{J})$ an $\bar{p} \in \mathcal{P}(\mathcal{I})$ exists such that $f(\bar{p}, \bar{q}) \in C$.

The assumption of Theorem 8.3 is satisfied for the set C, since for any mixed strategy $\bar{q} \in \mathcal{P}(\mathcal{J}) = \mathcal{P}(A)$ of the second player a mixed strategy $\bar{s}_k \in \mathcal{P}_\epsilon$ exists such that $\|\bar{s}_i - \bar{q}\| \leqslant \epsilon$, ie, $f(k, \bar{q}) \in C$. This probability distribution $\bar{s}_k$ is the mixed strategy of the first player.

[2] The condition (8.19) is equivalent to (8.20).

[3] We identify the set $\mathcal{P}_\epsilon = \{\bar{s}_1, \ldots, \bar{s}_N\}$ with the set of indices $\mathcal{I} = \{1, 2, \ldots, N\}$ of its elements.

In Theorem 8.2, we take $\bar{p}$ be equal to the pure strategy $\bar{\delta}[k]$ on $\mathcal{I} = \{1, \dots N\}$ which is concentrated in the number k, where $1 \leqslant k \leqslant N$.

By Theorem 8.2 a randomized strategy $P = \prod P_t$ of *Forecaster* exists, where $P_t \in \mathcal{P}(\mathcal{I})$, such that, regardless of the *Nature* moves $a_1, a_2, \dots$, the sequence of vector valued gains

$$\frac{1}{T}\sum_{t=1}^{T} f(i_t, a_t) = \begin{pmatrix} \frac{1}{T}\sum\limits_{t=1}^{T} I_{\{i_t=1\}}(\bar{s}_1 - \bar{\delta}[a_t]) \\ \dots \\ \frac{1}{T}\sum\limits_{t=1}^{T} I_{\{i_t=N\}}(\bar{s}_N - \bar{\delta}[a_t]) \end{pmatrix}.$$

P-almost surely approaches to the set C, where the trajectory $i_1, i_2, \dots$ is distributed by the measure P.

Hence, the condition (8.20) of calibration holds almost surely.

Theorem is proved. $\triangle$

The sequence of forecasts is said to be well-calibrated for a sequence of outcomes if it is ϵ-calibrated for each $\epsilon > 0$.

Predictions, which are chosen from a finite set $\mathcal{P}_\epsilon = \{\bar{s}_1, \ \dots, \ \bar{s}_{N_\epsilon}\}$ and satisfy the condition (8.20), are called ϵ-calibrated predictions.

We can strengthen Theorem 8.4 and obtain well-calibrated predictions.

Theorem 8.5. *A randomized strategy P of Forecaster can be constructed such that for each $\bar{p} \in \mathcal{P}(A)$, the condition of calibration*

$$\lim_{T\to\infty} \left\| \frac{1}{T}\sum_{t=1}^{T} I_{\|\bar{p}_t - \bar{p}\| \leqslant \epsilon}(\bar{p}_t - \bar{\delta}[a_t]) \right\| = 0 \tag{8.21}$$

holds P-almost surely, where the sequence $\bar{p}_1, \bar{p}_2, \dots$ is distributed by the measure P.

Scetch of the proof. Let ϵ_i be a sequence of rational numbers such that $\epsilon_i \to 0$ as $i \to \infty$. To construct the required sequence of predictions it is necessary to divide all rounds of the construction on sessions of sufficiently large size. For each such session, for each i, we define forecasts which are ϵ_i-calibrated at the right end-point of the

series of the session, and are ϵ_{i-1}-calibrated at the left end-point of this session. We omit details of this construction.

Using the same construction, we can strengthen (8.20) to

Theorem 8.6. *A randomized strategy of Forecaster can be constructed such that*

$$\lim_{T\to\infty} \sum_{\bar{p}\in\mathcal{P}(A)} \left\| \frac{1}{T} \sum_{t=1}^{T} I_{\{\bar{p}_t=\bar{p}\}} (\bar{\delta}[a_t] - \bar{p}) \right\| = 0 \tag{8.22}$$

holds almost surely.

Note that in the first sum (8.22) only a finite number od addends are nonzero: the summing is only by $\bar{p} \in \{\bar{p}_t : 1 \leqslant t \leqslant T\}$.

We pass details of the proof.

8.4. Calibrated forecasting and correlated equilibrium

In this section we show that if in some infinitely repeated game each player uses the predictions of future moves of opponents which are well-calibrated on a sequence of strategies chosen by these opponents, and chooses "the best reply" to these predictions, the joint frequency distribution of the players' strategies converges to the set of correlated equilibria of the game.

Any probability distribution in a finite set of cardinality N is an N-dimensional vector $\bar{p}$. We use a norm $\|\bar{p}\|$ on $\mathcal{R}^N$ and the corresponding distance $\operatorname{dist}(\bar{p},\bar{q}) = \|\bar{p}-\bar{q}\|$. Since all such norms are equivalent, it is not important which norm we use.

A distance from any element $\bar{p} \in \mathcal{R}^N$ to a set $S \subseteq \mathcal{R}^N$ is defined:

$$\operatorname{dist}(\bar{p}, S) = \inf_{\bar{q}\in S} \operatorname{dist}(\bar{p},\bar{q}).$$

An infinite sequence $\bar{p}_1, \bar{p}_2, \dots$ converges to a set S if

$$\lim_{t\to\infty} \operatorname{dist}(\bar{p}_t, S) = 0.$$

Consider a game with k players, where $A_i = \{1, \ldots, N_i\}$ is a set of all strategies of a player i, $i = 1, \ldots, k$. Also, for any i, let $f^i(i_1, \ldots, i_k)$ be a payoff function of player i, where $i_s \in A_s$, $s = 1, \ldots k$ are moves of all players.

A mixed strategy of a player s is a probability distribution in the set A_s of his strategies. We also consider joint mixed strategies of ordered sets $s_1, \ldots, s_l$ of players that are joint probability distributions in the sets $A_{s_1} \times \cdots \times A_{s_l}$ of ordered sets of their strategies.

Let $A = \prod_{j=1}^{k} A_j$ and $A_{-i} = \prod_{j \neq i} A_j$. Let also, $\bar{p}^t_{-i}$ be an arbitrary probability distribution in the set of strategies of all players excluding i. Here the lower index "$-i$" emphasizes that $\bar{p}_{-i} \in \mathcal{P}(A_{-i})$.

We alse use notations:

$$f^i(a, \bar{p}_{-i}) = E_{\bar{p}_{-i}}(f^i(a, \cdot)) = \sum_{\bar{a}_{-i} \in A_{-i}} f^i(a, \bar{a}_{-i}) \bar{p}_{-i}(\bar{a}_{-i}),$$

$$\bar{a}_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_k),$$

$$(a, \bar{a}_{-i}) = (a_1, \ldots, a_{i-1}, a, a_{i+1}, \ldots, a_k),$$

where $a \in A_i$, $\bar{a}_{-i} \in A_{-i}$, $E_{\bar{p}_{-i}}$ is a symbol of the mathematical expectation with respect to the measure $\bar{p}_{-i}$.

Now let players repeat the game at steps $t = 1, 2, \ldots$ according to the following protocol.

FOR $t = 1, 2, \ldots$

For any $i = 1, \ldots, k$, the player i outputs a forecast of a set of prediction of future moves of its opponents $j \neq i$ that is a probability distribution $\bar{p}^t_{-i}$ (the joint mixed strategy of all players $j \neq i$) and chooses the strategy $a^t_i \in A_i$ such that the payoff of player i is maximal provided that his opponents will choose the joint mixed strategy $\bar{p}^t_{-i}$:

$$a^t_i \in \operatorname{argmax}_{a \in A_i} f^i(a, \bar{p}^t_{-i}). \tag{8.23}$$

ENDFOR

We call a strategy a of the player i *the best response* for the forecast $\bar{p}^t_{-i}$ of moves of players $j \neq i$ if the value $\bar{f}^i(a, \bar{p}^t_{-i})$ is maximal:

$$\bar{f}^i(a, \bar{p}^t_{-i}) = \max_x \bar{f}^i(x, \bar{p}^t_{-i}).$$

If there are several of these strategies, we choose one of such $a = a_i^t$ using any pre-fixed rule.

Let $\bar{a}^t = (a_1^t, \ldots, a_k^t)$ be an ordered set of moves of all players at step t. Let

$$\bar{p}_T = \frac{1}{T} \sum_{t=1}^{T} \bar{\delta}[\bar{a}^t] \tag{8.24}$$

be the empirical frequency distribution of all strategies choosen over first T rounds of the game. Here $\bar{\delta}[\bar{a}]$ is a vector of dimension $\prod_{i=1}^{k} |A_i|$, where a coordinate corresponding to a vector $\bar{a}$ is 1, and all other coordinates are 0.

The coordinates of the vectors $\bar{p}_T$ are frequencies of occuring of each ordered set of strategies $\bar{a} = (a_1, \ldots, a_k)$ in the sequence of all ordered sets $\bar{a}^t = (a_1^t, \ldots, a_k^t)$ chosen by players on steps $t = 1, \ldots, T$.

The dimension of the vector $\bar{p}_T$, and of the vector $\bar{\delta}[\bar{a}^t]$, is equal to the number of all ordered sets $(a_1^t, \ldots, a_k^t)$, ie, to the number $\prod_{i=1}^{k} |A_i|$.

Any ordered set of strategies $\bar{a} = (a_1, \ldots, a_k)$ defines a number

$$\bar{p}_T(\bar{a}) = \frac{1}{T} |\{t : 1 \leqslant t \leqslant T, \bar{a}^t = \bar{a}\}| \tag{8.25}$$

that is a frequency of occurring the vector $\bar{a}$ in the sequence of ordered sets of strategies $\bar{a}^1, \ldots, \bar{a}^T$.

The following theorem shows that if each player

- uses the predictions of moves of all other players, which are well-calibrated in sense of (8.22) on a sequence of ordered sets of strategies choosen by his opponents, and
- chooses the best response (8.23) for these predictions,

then the joint frequency distribution of the players' strategies converges to the set $\mathcal{C}$ of correlated equilibria of the game.

Theorem 8.7. *Let for each i the sequence of forecasts $\bar{p}_{-i}^1, \bar{p}_{-i}^2, \ldots$ of the player i is well-calibrated for a sequence $\bar{a}_{-i}^1, \bar{a}_{-i}^2, \ldots$ of moves*

of all his opponents. Then the sequence of empirical frequency distributions $\bar{p}_T$ defined by (8.24) converges to a set $\mathcal{C}$ of the correlated equilibria.

Proof. To prove this theorem we have to show that

$$\operatorname{dist}(\bar{p}_T, \mathcal{C}) \to 0$$

as $T \to \infty$, where $\mathcal{C}$ is the set of correlated equilibria. We also will prove that $\mathcal{C} \neq \emptyset$.

The simplex of all probability distributions in a polyhedron $A = \prod_{i=1}^{k} A_i$ (vectors of dimension $|A|$) is a compact set. Therefore, the sequence of frequency distributions $\{\bar{p}_T : T = 1, 2, \dots\}$ defined by (8.24) contains an infinite convergent subsequence $\bar{p}_{T_j}$.

Let $\bar{p}^*$ be a limit point of this subsequence. We prove that $\bar{p}^*$ is a correlated equilibrium.

Fix an i and a strategy $a \in A_i$ of the player i such that

$$\bar{p}^*(a) = \sum_{\bar{a}: a_i = a} \bar{p}^*(\bar{a}) > 0,$$

where $\bar{a} = (a_1, \dots, a_k)$, $a_j \in A_j$, $j = 1, \dots, k$. [4]

We write $f = f^i$, and define two subsets $B, \tilde{B} \subseteq \mathcal{P}(A_{-i})$ (depending on i and a):

$$B = \{\bar{q}_{-i} : \bar{f}(a, \bar{q}_{-i}) = \max_{a' \in A_i} \bar{f}(a', \bar{q}_{-i})\}$$

be a set of all mixed strategies of all opponents of the player i, for which its pure strategy a is the best response. It is easy to see that B is a closed convex set. We also define

$$\tilde{B} = \left\{ \bar{q}_{-i} : \exists t \left(\bar{q}_{-i} = \bar{q}^t_{-i}) \& \bar{f}(a, \bar{q}_{-i}) = \max_{a' \in A_i} \bar{f}(a', \bar{q}_{-i}) \right) \right\}$$

be a set of all mixed strategies chosen by the opponents of the player i at steps $t = 1, 2, \dots$, where he chooses the move a as the best response. By definition $\tilde{B} \subseteq B$.

[4] If $\bar{p}^*(a) = 0$ then the strategy a can be ignored when calculating the frequency distribution. This is equivalent to the case where the ith player does not use a.

By definition the set $\tilde{B}$ is no more than countable, since at each step no more than one element can be added to it.

Let us consider the conditional probability of an arbitrary vector of moves $\bar{a}_{-i}$ of all players, except i, given $a_i = a$ (where a has been chosen above) with respect to the limit distribution $\bar{p}^*$:

$$\bar{p}^*(a_{-i}|a_i = a) = \bar{p}^*((a, \bar{a}_{-i})|a_i = a) = \frac{\bar{p}^*(a, \bar{a}_{-i})}{\bar{p}^*(a_i = a)}. \tag{8.26}$$

By Corollary 6.3 a probability distribution $\bar{p}$ in the set $\prod_{k=1}^{K} A_k$ of sequences of moves $\bar{a} = (a_1, \ \ldots, \ a_K)$ is a correlated equilibrium if and only if for each player $i \in \{1, \ \ldots, \ K\}$ and for each strategy $a \in A_i = \{1, \ \ldots, \ N_i\}$

$$f^i(a, \bar{p}(\cdot|a_i = a)) = \max_{a' \in A_i} f^i(a', \bar{p}(\cdot|a_i = a)).$$

Then the probability distribution $\bar{p}^*$ is correlated equilibrium if and only if the conditional distribution $\bar{p}^*(\cdot|a_i = a) \in B$ for all i and $a \in A_i$.

We prove that $\bar{p}^*(\cdot|a_i = a) \in B$ by approximating it using the corresponding frequency distribution. Let for any $a \in A_i$,

$$N_T(a) = |\{t : 1 \leqslant t \leqslant T, \ a_i^t = a\}|$$

be the number of steps $\leqslant T$, on which the player i chooses the strategy a, and

$$N_T(\bar{p}_{-i}) = |\{t : 1 \leqslant t \leqslant T, \ \bar{p}_{-i}^t = \bar{p}_{-i}\}|$$

be the number of steps $\leqslant T$, on which the opponents of the player i choose an ordered set of mixed strategies $\bar{p}_{-i} \in \mathcal{P}(A_{-i})$.

Consider the conditional frequency distribution $\bar{p}_T(\cdot|a_i = a)$ of $\bar{a}_{-i}$ given $a_i = a$. This distribution is defined:

$$\bar{p}_T(a_{-i}|a_i = a) = \frac{\bar{p}_T(a, \bar{a}_{-i})}{\bar{p}_T(a_i = a)}. \tag{8.27}$$

By (8.26), $\bar{p}_{T_j}(a_{-i}|a_i = a) \to p^*(a_{-i}|a_i = a)$ as $j \to \infty$.

By definition of the set $\tilde{B}$, an element $a \in A_i$ appears in the set of strategies $\bar{a}^t$ as the ith coordinate only if $\bar{p}^t_{-i} \in \tilde{B}$. It follows that the frequency of occurrence of any set $(a, \bar{a}_{-i})$ in the sequence

$$\{\bar{a}^t : \ 1 \leqslant t \leqslant T\}$$

equal to the frequency of occurrence of an ordered set $\bar{a}_{-i}$ in the sequence

$$\{\bar{a}^t_{-i} : \bar{p}^t_{-i} \in \tilde{B}, \ 1 \leqslant t \leqslant T\}.$$

Then by (8.25) we obtain:

$$\bar{p}_T(a, \bar{a}_{-i}) = \bar{p}_T(\bar{a}) = \frac{1}{T}|\{t : 1 \leqslant t \leqslant T, \ \bar{p}^t_{-i} \in \tilde{B}, \ \bar{a}^t_{-i} = \bar{a}_{-i}\}|.$$

By definition

$$\bar{p}_T(a_i = a) = \frac{N_T(a)}{T}.$$

Hence, we obtain an expression for the conditional frequency distribution defined by a sequence $\bar{a}^t$, where $a^t_i = a$, $t = 1, \ldots, \ T$:

$$\bar{p}_T(\cdot|a_i = a) = \frac{1}{N_T(a)} \sum_{1 \leqslant t \leqslant T, \ \bar{p}^t_{-i} \in \tilde{B}} \bar{\delta}[\bar{a}^t_{-i}] =$$

$$= \left(\frac{T}{N_T(a)}\right) \frac{1}{T} \sum_{1 \leqslant t \leqslant T, \ \bar{p}^t_{-i} \in \tilde{B}} (\bar{\delta}[\bar{a}^t_{-i}] - \bar{p}^t_{-i}) + \tag{8.28}$$

$$+ \sum_{1 \leqslant t \leqslant T, \ \bar{p}_{-i} \in \tilde{B}} \left(\frac{N_T(\bar{p}_{-i})}{N_T(a)}\right) \bar{p}_{-i}. \tag{8.29}$$

Since $\bar{p}^*(a_i = a) > 0$ and $\bar{p}^*$ is the limit of probability distributions $\bar{p}_{T_j}$ as $j \to \infty$, then the factor $\left(\frac{T_j}{N_{T_j}(a)}\right)$ is bounded from above.

Since $\bar{p}_{T_j} \to \bar{p}^*$ as $j \to \infty$, then for any vector $\bar{a}_{-i}$,

$$\bar{p}_{T_j}(\bar{a}_{-i}|a_i = a) \to \bar{p}^*(a_{-i}|a_i = a)$$

as $T_j \to \infty$. Then, since the set B is closed, we obtain that $\bar{p}^*(\cdot|a_i = a) \in B$ for all i and all $a \in A_i$. Therefore, we have proved that the

probability distribution $\bar{p}^*$ is a correlated equilibrium. From this the theorem follows. $\triangle$

Combining Theorems 8.4 and 8.7, we obtain the following corollary:

Corollary 8.1. *One can construct a randomized algorithm which, for each $1 \leqslant i \leqslant N$ computes a sequence of predictions $\bar{p}^1_{-i}, \bar{p}^2_{-i}, \dots$ of the player's i opponents moves such that the following holds:*

- *Let each player chooses his move as the best response to a prediction of this algorithm.*
- *Then the empirical frequencies $\bar{p}_T$ of all players moves converge to a set $\mathcal{C}$ of correlated equilibria as $T \to \infty$ with probability 1.*

8.5. Problems

1. Prove the inequality (8.9).

2. Prove that the minimax theorem is the corollary of the Blackwell approachability theorem.

3. Prove that if a closed set S is approachable in a game with a matrix $f(i,j)$ then any closed subset of its complement cannot be approachable in the game with the matrix $f'(i,j) = f(j,i)$.

Part IV

Appendix

8.6. Some remarkable inequalities

In this section we give some large deviation inequalities that are repeatedly used in proofs of theorems. The most important is that the Hoeffding inequality.

Lemma 8.1. *Let X be a random variable such that $a \leqslant X \leqslant b$, where a and b are some real numbers, $a < b$. Then for any real number s,*

$$\ln E(e^{sX}) \leqslant sE(X) + \frac{s^2(b-a)^2}{8}, \tag{8.30}$$

where E is a symbol of the mathematical expectation.

Proof. Since

$$\ln E(e^{sX}) = sE(X) + \ln E(e^{s(X-E(X))}),$$

it is sufficient to prove that for any random variable X such that $E(X) = 0$ and $a \leqslant X \leqslant b$

$$E(e^{sX}) \leqslant e^{s^2(b-a)^2/8}.$$

By convexity of the exponent

$$e^{sx} \leqslant \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

for $a \leqslant x \leqslant b$.

Denote $p = -\frac{a}{b-a}$. Applying the mathematical expectation to both sides of this inequality and taking into account that $E(X) = 0$, we obtain for $x = X$:

$$E(e^{sX}) \leqslant -\frac{a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa} =$$
$$= (1 - p + pe^{s(b-a)})e^{-ps(b-a)} = e^{\varphi(u)},$$

where $u = s(b-a)$ and $\varphi(u) = -pu + \ln(1 - p + pe^u)$.

The first derivative of $\varphi(u)$ by u is represented as

$$\varphi'(u) = -p + \frac{p}{p + (1-p)e^{-u}}.$$

It holds $\varphi(0) = \varphi'(0) = 0$. Besides,

$$\varphi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leqslant \frac{1}{4}.$$

Indeed, denote $q = (1-p)e^{-u}$. We need to prove the inequality $\frac{pq}{(p+q)^2} \leqslant \frac{1}{4}$, which follows from $(p-q)^2 \geqslant 0$.

By Taylor's formula, for some $\theta \in [0, u]$,

$$\varphi(u) = \varphi(0) + u\varphi'(0) + \frac{u^2}{2}\varphi''(\theta) \leqslant \frac{u^2}{8} = \frac{s^2(b-a)^2}{8},$$

since $u = s(b-a)$. Lemma is proved. $\triangle$

Let us consider several corollaries, explaining the importance of this inequality.

Corollary 8.2. *Let X be a random variable such that $P\{a \leqslant X \leqslant b\} = 1$. Then*

$$P\{|X - E(X)| > c\} \leqslant 2e^{-\frac{2c^2}{(b-a)^2}}. \tag{8.31}$$

Proof. First recall the Markov inequality. Let X be a random variable, $X \geqslant 0$. It follows from

$$E(X) = \int X dP \geqslant \int_{\{X>c\}} X dP \geqslant cP\{X > c\}$$

that $P\{X > c\} \leqslant E(X)/c$.

Using Markov inequality and the inequality (8.30), we obtain

$$P\{X - E(X) > c\} = P\{e^{s(X-E(X))} > e^{cs}\} \leqslant e^{-cs+\frac{s^2(b-a)^2}{8}}$$

for all $s > 0$. The minimum of the right-hand side of this inequality by s is attained for $s = 4c/(b-a)^2$. From this we obtain

$$P\{X - E(X) > c\} \leqslant e^{-\frac{2c^2}{(b-a)^2}}.$$

Similarly, we get

$$P\{X - E(X) < -c\} \leqslant e^{-\frac{2c^2}{(b-a)^2}}.$$

Finally, we obtain

$$P\{|X - E(X)| > c\} \leqslant 2e^{-\frac{2c^2}{(b-a)^2}}.$$

$\triangle$

The best known is following corollary from this lemma – Chernoff inequality. [5]

Corollary 8.3. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables such that $P\{a_i \leqslant X \leqslant b_i\} = 1$ for all $i = 1, 2, \ldots$. Then for any $c > 0$,*

$$P\left\{\sum_{i=1}^{n} X_i - E\sum_{i=1}^{n} X_i > c\right\} \leqslant \exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right),$$

and also,

$$P\left\{\sum_{i=1}^{n} X_i - E\sum_{i=1}^{n} X_i < -c\right\} \leqslant \exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right).$$

Proof. The proof is similar to the proof of Corollary 8.2. By

[5] We use also notation $\exp(x) = e^x$.

Markov inequality and by the inequality (8.30) we obtain

$$P\left\{\sum_{i=1}^{n}(X_i - E(X_i)) > c\right\} \leqslant$$

$$\leqslant \frac{E\left(\exp(s\sum_{i=1}^{n}(X_i - E(X_i)))\right)}{\exp(cs)} =$$

$$= \frac{\prod_{i=1}^{n} E(\exp(s(X_i - E(X_i))))}{\exp(cs)} \leqslant$$

$$\leqslant \frac{\prod_{i=1}^{n} \exp\left(\frac{s^2(b_i-a_i)^2}{8}\right)}{\exp(cs)} \leqslant$$

$$\leqslant \exp\left(-cs + \frac{s^2\sum_{i=1}^{n}(b_i - a_i)^2}{8}\right) \leqslant$$

$$\leqslant \exp\left(-\frac{2c^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

In transition from the second line to the third line we have used the independence of the random variables $X_1, X_2, \ldots$. In the transition from the penultimate line to the last line we have used the minimization by s. The second inequality is obtained similarly. $\triangle$

Using this corollary, it is possible to obtain a bound for the rate of convergence for the law of of large numbers.

Corollary 8.4. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables such that $P\{a_i \leqslant X \leqslant b_i\} = 1$ for all $i = 1, 2, \ldots$. Then*

for any $\epsilon > 0$,

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i))\right| > \epsilon\right\} \leqslant 2\exp\left(-\frac{2n^2\epsilon^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right).$$

If $a_i = 0$ *and* $b_i = 1$ *for all* i *then*

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i))\right| > \epsilon\right\} \leqslant 2\exp\left(-2n\epsilon^2\right).$$

A sequence of random variables $V_1, V_2, \ldots$ is called a martingale-difference relative to the sequence of random variables $X_1, X_2, \ldots$ if for each $i > 1$ the variable V_i is a function of random variables $X_1, \ldots, X_i$ and

$$E(V_{i+1}|X_1, \ldots, X_i) = 0$$

with probability one. The following inequality is called Azuma–Hoeffding inequality.

Lemma 8.2. *Let* $V_1, V_2, \ldots$ *be a martingale-difference relative a sequence* $X_1, X_2, \ldots$ *of random variables, besides,* $V_i \in [A_i, A_i + c_i]$ *for some random variable* A_i *measurable with respect to* $X_1, \ldots, X_{i-1}$ *and for some sequence of positive constants* $c_1, c_2, \ldots$.

For $S_n = \sum\limits_{i=1}^{n} V_i$, *it holds*

$$E(e^{sS_n}) \leqslant e^{(s^2/8)\sum\limits_{i=1}^{n} c_i^2}$$

for all $s > 0$.

Proof. We have

$$\begin{aligned} E(e^{sS_n}) = E(e^{sS_{n-1}}E(e^{sV_n}|X_1, \ldots, X_{n-1})) \leqslant \\ \leqslant E(e^{sS_{n-1}}e^{s^2c_n^2/8}) = \\ = e^{s^2c_n^2/8}E(e^{sS_{n-1}}). \end{aligned} \tag{8.32}$$

Here in transition from the first line to the second line Lemma 8.1 was used. The result of the lemma can be obtained by iteration of the inequality (8.32). $\triangle$

The following corollary is proved similarly to Corollary 8.2.

Corollary 8.5. *Let $V_1, V_2, \ldots$ be a martingale-difference relative a sequence $X_1, X_2, \ldots$ of random variables, besides, $V_i \in [A_i, A_i + c_i]$ for some random variable A_i measurable with respect to $X_1, \ldots, X_{i-1}$ and for some sequence of positive constants $c_1, c_2, \ldots$.*

Denote $S_n = \sum\limits_{i=1}^{n} V_i$. Then for any $n > 0$,

$$P\{|S_n| > c\} \leqslant 2 \exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n} c_i^2} \right).$$

Proof. Using Markov inequality

$$P\{X > c\} \leqslant E(X)/c$$

and the inequality (8.30), we obtain for any n,

$$P\{S_n > c\} = P\{e^{sS_n} > e^{cs}\} \leqslant \exp\left(-cs + \frac{s^2 \sum\limits_{i=1}^{n} c_i^2}{8} \right)$$

for all $s > 0$. The minimum of the right-hand side by s is attained for $s = 4c/\sum\limits_{i=1}^{n} c_i^2$. From this we obtain

$$P\{S_n > c\} \leqslant \exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n} c_i^2} \right). \tag{8.33}$$

Similarly,

$$P\{S_n < -c\} \leqslant \exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n} c_i^2} \right).$$

Finally, we obtain

$$P\{|S_n| > c\} \leqslant 2\exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n} c_i^2}\right).$$

Corollary 8.6. *Under the conditions of Corollary 8.5, where $c_i = 1$ for all i, we have*

$$P\left\{\frac{1}{n}|S_n| > c\right\} \leqslant 2e^{-2nc^2}. \tag{8.34}$$

Borel–Cantelli lemma states that if, for some sequence of events A_n the series $\sum\limits_{n=1}^{\infty} P(A_n)$ converges, then the probability that the event A_n holds for infinitely many n is 0.

Since for any $c > 0$ the series of exponents in the right-hand side of the inequality (8.34) is convergent, by Borel–Cantelli lemma we obtain:

Corollary 8.7. *Under the conditions of Corollary 8.5, where $B_1 < c_i < B_2$ for all i, for some positive constants B_1 and B_2, the martingale strong law of large numbers holds:*

$$P\left\{\lim_{n\to\infty} \frac{S_n}{n} = 0\right\} = 1. \tag{8.35}$$

Bibliography

[1] Alon N., Ben-David S., Cesa-Bianchi N., Haussler D. Scale-sensitive dimensions, uniform convergence, and learnability. J. ACM V. 1977. 44(4). P. 615-631.

[2] Aronszajn N. Theory of reproducing kernels. Transactions of the American Mathematical Society. 1950. V. 68. P. 337404.

[3] Bartlett P., Mendelson S. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. Journal of Machine Learning Research. 2002. V.3. P. 463-482.

[4] Bartlett P., Bousquet O., Mendelson S. Local Rademacher Complexities. The Annals of Statistics. 2005, V. 33, No. 4, 14971537.

[5] Beckenbach E.F., Bellman R. Inequalities (Mathematics). Berlin[Germany: West]: Springer–Verlag, 1971.

[6] Blackwell D. An analog of the minimax theorem for vector payoffs. Pacific Journal of Mathematics. 1956. V. 6. P. 1–8.

[7] Bousquet, Olivier, Stephane Boucheron, and Gabor Lugosi. Introduction to statistical learning theory. Advanced Lectures on Machine Learning. 2004. P. 169-207.

[8] A. Chernov, F. Zhdanov. Prediction with expert advice under discounted loss. Technical report, arXiv:1005.1918v1 [cs.LG], 2010.

[9] Cover T., Ordentlich E. Universal portfolio with side information. IEEE Transaction on Information Theory – 1996. – V. 42. – P. 348–363.

[10] Cristianini N., Shawe-Taylor J. An Introduction to Support Vector Machines. – Cambridge UK: Cambridge University Press, 2000.

[11] A.P. Dawid. The well-calibrated Bayesian [with discussion]. J. Am. Statist. Assoc. 77, 1982, 605–613

[12] Dawid A.P. Calibration-based empirical probability [with discussion]. Ann. Statist. – 1985. – V. 13. – P. 1251–1285.

[13] Foster D.P., Vohra R. Asymptotic calibration. Biometrika. – 1998. – V. 85. – P. 379–390.

[14] Freund Y., Schapire R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences – 1997. – V. 55. – P. 119–139.

[15] J. Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A.W. Tucker, and P. Wolfe, editors, Contributions to the Theory of Games 3, pages 97-139, Princeton University Press, 1957.

[16] M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader. Journal of Machine Learning Research, 6:639–660, 2005.

[17] Kakade, S.M., Foster, D.P. Deterministic calibration and Nash equilibrium. Lecture Notes in Computer Science – Berlin: Springer, 2004. – V. 3120. – P. 33–48.

[18] Sham Kakade and Ambuj Tewari. Topics in Artificial Intelligence (Learning Theory) - Spring 2008. Lecture Notes. *http* : *//ttic.uchicago.edu/ tewari/LT*$_S$*P*2008*.html*

[19] A. Kalai and S. Vempala. Efficient algorithms for online decisions. In Bernhard Scholkopf, Manfred K. Warmuth, editors,

Proceedings of the 16th Annual Conference on Learning Theory COLT 2003, Lecture Notes in Computer Science 2777, pages 506–521, Springer-Verlag, Berlin, 2003. Extended version in Journal of Computer and System Sciences, 71:291–307, 2005.

[20] Kimeldorf G. S. and Wahba G. Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. – 1971 –V. 33 – 8295.

[21] Ledoux, M., Talagrand, M. Probability in Banach Spaces: Isoperimetry and Processes. Springer, New York. 1991.

[22] Littlestone N., Warmuth M. The weighted majority algorithm. Information and Computation – 1994 – V. 108 – P. 212–261.

[23] Lugosi G., Cesa-Bianchi N. Prediction, Learning and Games. – New York: Cambridge University Press, 2006.

[24] Mannor S., Stoltz G. A Geometric Proof of Calibration. arXiv:0912.3604v2. 2009.

[25] McDiarmid C. On the method of bounded differences. London Mathematical Society Lecture Notes Series. Surveys in Combinatorics. Cambridge University Press. V. 141. pp. 148188. 1989.

[26] Shafer G., Vovk V. Probability and Finance. It's Only a Game! – New York: Wiley. 2001.

[27] Shawe-Taylor J., Cristianini N. Margin distribution bounds on generalization. In Proceedings of the European Conference on Computational Learning Theory, EuroCOLT'99. P.263–273. 1999.

[28] Shawe-Taylor J., Cristianini N. Kernel Methods for Pattern Analysis. – Cambridge UK: Cambridge University Press, 2004.

[29] Shiryaev A.N. Probability. Springer-Verlag, Berlin, 1980

[30] Scholkopf B. and Smola A. Learning with Kernels. MIT Press, Cambridge, MA, 2002.

[31] Steinwart I. On the influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research, 2, 67-93, 2001

[32] Valiant L.G. A theory of the learnable, Communications of the ACM V. 27(11). P. 1134-1142. 1984.

[33] Vapnik, V., Chervonenkis, A. Theory of Pattern Recognition. Nauka, Moscow (1974) (in Russian); German translation: Theorie der Zeichenerkennung, Akademie Verlag, Berlin, 1979.

[34] Vapnik V.N. Statistical Learning Theory. – New York: Wiley, 1998.

[35] Vovk V. Aggregating strategies. Proceedings of the 3rd Annual Workshop on Computational Learning Theory (M. Fulk and J. Case, editors,) – San Mateo, CA: Morgan Kaufmann, 1990. – P. 371–383.

[36] Vovk V. A game of prediction with expert advice. Journal of Computer and System Sciences – 1998 – V. 56. – No. 2.
P. 153–173.

[37] Vovk V., Watkins C. Universal portfolio selection. Proceedings of the 11th Annual Conference on Computational Learning Theory – New York: ACM Press, 1998. – P. 12–23.

[38] Vovk V. Competitive on-line statistics. International Statistical Review – 2001 – V. 69. – P. 213–248.

[39] Vovk V, Gammerman A., Shafer G. Algorithmic Learning in a Random World. Springer, New York, 2005.

[40] Vovk V., Shafer G. Good randomized sequential probability forecasting is always possible. J. Royal Stat. Soc. B. – 2005 – V. 67 – P. 747–763.

[41] V. Vovk. On-line regression competitive with reproducing kernel Hilbert spaces (extended abstract). TAMS Lecture Notes in Computer Science – Berlin: Springer, 3959, 2006, 452–463

[42] Vovk V., Takemura A., Shafer G. Defensive forecasting. Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (ed. by R. G. Cowell and Z. Ghahramani) – Cambridge UK: Society for Artificial Intelligence and Statistics, 2005. – P. 365–372.

[43] Vovk V. On-line regression competitive with reproducing kernel Hilbert spaces (extended abstract). Lecture Notes in Computer Science – Berlin: Springer, 2006. V. 3959. P. 452–463.

[44] Vovk V. Predictions as statements and decisions. Theoretical Computer Science 2008. V. 405. No. 3. P. 285–296.

[45] V'yugin V., Trunov V. Universal algorithmic trading. Journal of Investment Strategies. V.2 (1), Winter 2012/13 P. 63-88.