Int. J. Appl. Math. Comput. Sci., 2010, Vol. 20, No. 3, 525–544
DOI: 10.2478/v10006-010-0039-x

PROBABILITIES OF DISCREPANCY BETWEEN MINIMA OF CROSS–VALIDATION, VAPNIK BOUNDS AND TRUE RISKS

PRZEMYSŁAW KLĘSK

Department of Methods of Artificial Intelligence and Applied Mathematics
Westpomeranian University of Technology, ul. Żołnierska 49, 71–210 Szczecin, Poland
e-mail: pklesk@wi.zut.edu.pl

Two known approaches to complexity selection are taken under consideration: n-fold cross-validation and structural risk minimization. Obviously, in either approach, a discrepancy between the indicated optimal complexity (indicated as the minimum of a generalization error estimate or a bound) and the genuine minimum of *unknown true risks* is possible. In the paper, this problem is posed in a novel quantitative way. We state and prove theorems demonstrating how one can calculate pessimistic probabilities of discrepancy between these minima for given for given conditions of an experiment. The probabilities are calculated in terms of all relevant constants: the sample size, the number of cross-validation folds, the capacity of the set of approximating functions and bounds on this set. We report experiments carried out to validate the results.

Keywords: regression estimation, model comparison, complexity selection, cross-validation, generalization, statistical learning theory, generalization bounds, structural risk minimization.

1. Introduction and notation

Practitioners typically apply an n-fold cross-validation procedure to select the best complexity for a model, given a data set of a certain size (Hjorth, 1994; Efron and Tibshirani, 1993). Obviously, it is a time-consuming procedure. Sometimes, for sufficiently large problems, it may take days of computations to accomplish the task.

On the other hand, there is the structural risk minimization approach proposed by Vapnik as a part of his *statistical learning theory* (Vapnik 1995; 1998; 2006, Bousquet *et al.*, 2004). The approach is based on probabilistic bounds on the generalization of learning machines. The key mathematical tools applied to derive the bounds in their additive versions are Chernoff and Hoeffding inequalities[1] (Vapnik, 1998; Cherkassky and Mulier, 1998; Hellman and Raviv, 1970; Schmidt *et al.*, 1995). To select the best complexity for a model, one iterates over successive complexities and looks at the minimum point of *bounds* on generalization errors, instead of looking at estimates of these errors via cross-validation. Since the bound is calculated only *once* for a fixed complexity, the approach is $O(n)$ times faster than cross-validation. Yet, if the data set at our disposal is *small*[2], the minimum point indicated via SRM is usually underestimated, since a summand in the bound related to model complexity—the *capacity* of the set of functions—is strongly pessimistic (Vapnik, 1998; Anthony and Shawe-Taylor, 1993; Krzyżak et al., 2000; Shawe-Taylor *et al.*, 1996).

Although the name SRM tells it explicitly, clearly in both approaches—cross-validation and SRM—one iterates over the so-called *structure*, i.e., a sequence of nested sets of approximating functions, which constitutes an increasing complexity.

We remark that in both approaches the modeler is uncertain whether the complexity he/she chose as the point with the minimum generalization error estimate or bound is truly the minimum point of *unknown true risks* and therefore the genuine optimal complexity. An example of such a possible discrepancy between these three minima is shown in Fig. 1.

In the paper we state and prove theorems asserting how one can calculate probabilities of discrepancy betwe-

[1]Chernoff inequality is $P\big(|\nu_I - p| \geq \epsilon\big) \leq 2\exp(-2\epsilon^2 I)$, Hoeffding inequality is $P\big(|X_I - EX| \geq \epsilon\big) \leq 2\exp(-2\epsilon^2 I/(B^2 - A^2))$, meaning respectively that observed frequencies on a sample of size I converge to their true probabilities as I grows large. Analogically, the mean of a random variable (bounded by A and B) converges to its expected value. It is *in-probability-convergence* and its rate is exponential.

[2]Vapnik proposes to call a sample *small* if the ratio of its size to the Vapnik–Chervonenkis dimension is less than 20.

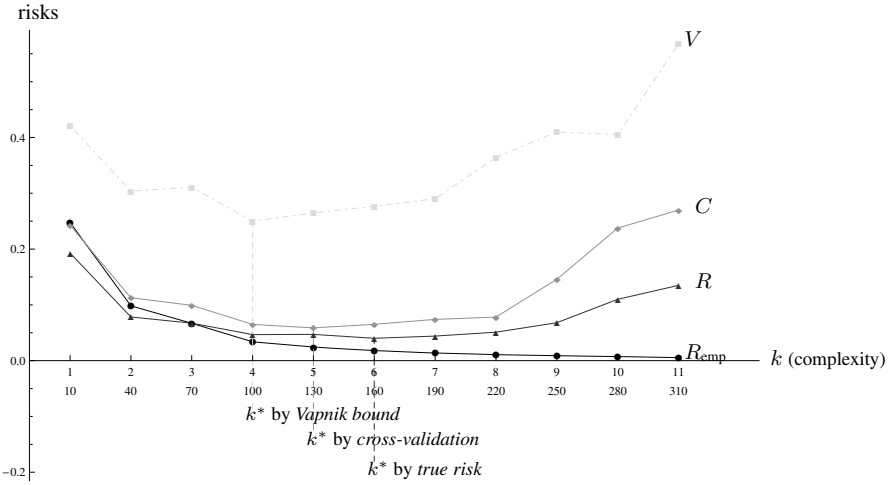

Fig. 1. Example illustration of the discrepancy between the minima k^* indicated by cross-validation, SRM (Vapnik bound) and true risk. On the horizontal axis, indices of complexities are shown $k \in \{1, 2, \ldots, 11\}$ (and numbers of terms in functions corresponding to them). On the vertical axis, the values of *risks* are shown: empirical risks R_{emp}, true risks R, cross-validation result C, Vapnik bounds V.

en minima of (a) cross-validation results, (b) Vapnik bounds, (c) true risks. We remark that while the values of (a) and (b) can be *known* (measured, calculated), the values of (c) are in practice *unknown*. In this sense, probabilities are interesting, because they assess discrepancy between something known and something that cannot be known.

The probabilities are calculated in terms of all relevant constants, such as the sample size, the number of cross-validation folds, the capacity of the set of approximating functions and bounds of this set.

According to the author's knowledge, this paper poses an original problem. Among works related to statistical learning and SRM we have not come across publications where the problem of calculating the *probabilities* of discrepancy between the above-mentioned minima was posed or taken up quantitatively. Latest works on the subject of generalization in machine learning follow rather different research directions like ϵ-covering numbers and fat-shattering dimension (Zhang, 2002; Bartlett *et al.*, 1997), regularization techniques (Hasterberg et al., 2008; Ng, 2004), or sample complexity (Bartlett, 1998; Bartlett and Tewari, 2007).

In the paper we focus on the *regression estimation* learning task, nevertheless the theorems and results can be broadened without difficulty also onto classification (pattern-recognition).

1.1. Notation related to statistical learning theory.

We use a notation similar to Vapnik's. We denote the finite set of samples as

$$\big\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_I, y_I)\big\},$$

or, more briefly, by encapsulating pairs as

$$\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_I\},$$

where $\mathbf{x}_i \in \mathbb{R}^d$ are input points and $y_i \in \mathbb{R}$ are output values corresponding to them[3].

We denote the *set of approximating functions* (models) by

$$\{f(\mathbf{x}, \omega)\}_{\omega \in \Omega},$$

where Ω is the domain of parameters of this set of functions, and a fixed ω can be regarded as an index of a specific function in the set.

The *risk functional* R: $\{f(\mathbf{x}, \omega)\}_{\omega \in \Omega} \to \mathbb{R}$ is defined as

$$R(\omega) = \int_{\mathbf{x} \in \mathbf{X}} \int_{y \in Y} L\Big(f(\mathbf{x}, \omega), y\Big) \underbrace{p(\mathbf{x}, y)}_{p(\mathbf{x})p(y|\mathbf{x})} \, \mathrm{d}y \, \mathrm{d}\mathbf{x}, \tag{1}$$

where $p(\mathbf{x})$ is a probability density of input $\mathbf{x}$, $p(y|\mathbf{x})$ is a conditional density of system/phenomenon outputs y given a fixed $\mathbf{x}$. $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$ is the joint density for pairs $(\mathbf{x}, y)$. In practice, $p(\mathbf{x}, y)$ is unknown but *fixed*, and hence we assume the pairs in the sample $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_I\}$ to be *i.i.d.*[4] (Bousquet et al., 2004; Cherkassky and Mulier, 1998; Devroye *et al.*, 1996; Vapnik, 1998).

[3] *Regression estimation* learning task.

[4] Independent, identically distributed.

L is the so-called *loss function* which measures the discrepancy between the output y and the model f. For regression estimation, L is usually chosen as the distance in L_2 metric:

$$L\Big(f(\mathbf{x},\omega),y\Big)=\big(f(\mathbf{x},\omega)-y\big)^2, \tag{2}$$

and then the risk functional becomes[5]

$$R(\omega)=\int_{\mathbf{x}\in\mathbf{X}}\int_{y\in Y}\big(f(\mathbf{x},\omega)-y\big)^2 p(\mathbf{x},y)\,\mathrm{d}y\,\mathrm{d}\mathbf{x}. \tag{4}$$

By ω_0 we denote the index of the best function $f(\mathbf{x},\omega_0)$ in the set, such that

$$R(\omega_0)=\inf_{\omega\in\Omega}R(\omega). \tag{5}$$

Since only a finite set of samples $\{\mathbf{z}_1,\ldots,\mathbf{z}_I\}$ is at our disposal, we cannot count on actually finding the best function $f(\mathbf{x},\omega_0)$. In fact, we look for its estimate with respect to the finite set of samples. We define the *empirical risk*:

$$R_{\text{emp}}(\omega)=\frac{1}{I}\sum_{i=1}^{I}L(y_i,f(\mathbf{x}_i,\omega)), \tag{6}$$

and by ω_I we denote the index of the function $f(\mathbf{x},\omega_I)$ such that

$$R_{\text{emp}}(\omega_I)=\inf_{\omega\in\Omega}R_{\text{emp}}(\omega) \tag{7}$$

(empirical risk minimization principle) (Vapnik and Chervonenkis, 1968; Vapnik and Chervonenkis, 1989; Cherkassky and Mulier, 1998).

For notational simplicity and further discussion, we introduce equivalent replacements:

$$(\mathbf{x},y)=\mathbf{z},$$
$$L\big(f(\mathbf{x},\omega),y\big)=Q\big(\mathbf{z},\omega\big).$$

In other words, instead of considering the set of approximating functions[6] $\{f(\mathbf{x},\omega)\}_{\omega\in\Omega}$, we equivalently consider the *set of error functions* $\{Q(\mathbf{z},\omega)\}_{\omega\in\Omega}$. It is a 1:1 correspondence[7]. Now, we write the true risk as

$$\begin{aligned}R(\omega)&=\int_{\mathbf{z}\in\mathbf{X}\times Y}Q(\mathbf{z},\omega)\underbrace{p(\mathbf{z})}_{p(\mathbf{x},y)}\,\mathrm{d}\mathbf{z}\\&=\int_{\mathbf{Z}}Q(\mathbf{z},\omega)\,\mathrm{d}F(\mathbf{z}),\end{aligned} \tag{8}$$

and the empirical risk as

$$R_{\text{emp}}(\omega)=\frac{1}{I}\sum_{i=1}^{I}Q(\mathbf{z}_i,\omega)). \tag{9}$$

1.2. Notation related to cross-validation. In the paper, we consider the *non-stratified* variant of the n-fold cross-validation procedure (Kohavi, 1995). In each single fold (iteration), we split the data set into two disjoint subsets—a training set and a testing set, but among folds we do not care that training sets themselves are *disjoint* pairwise. In other words, folds are independent. Such an approach is somewhere in-between the classical n-fold cross-validation and *bootstrapping* (Efron and Tibshirani, 1993). In the classical cross-validation, all $\binom{n}{2}$ pairs of training sets are mutually disjoint (and so are testing sets), whereas in bootstrapping, instead of repeatedly analyzing subsets of data, one repeatedly analyzes data subsamples (with replacement). For more information, see also the works of Hjorth (1994), Weiss and Kulikowski (1991) and Fu *et al.*(2005).

We introduce the following notation: I' and I'' stand for the sizes of training and testing sets

$$I'=\frac{n-1}{n}I,$$
$$I''=\frac{1}{n}I,$$

respectively. Without loss of generality for further theorems and proofs, let I be divisible by n, so that I' and I'' are integers.

In a single fold, let

$$\{\mathbf{z}'_1,\mathbf{z}'_2,\ldots,\mathbf{z}'_{I'}\},$$
$$\{\mathbf{z}''_1,\mathbf{z}''_2,\ldots,\mathbf{z}''_{I''}\}$$

represent respectively the training set and the testing set, taken as a random split of the whole data set $\{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_I\}$. Similarly, empirical risks calculated as follows:

$$R'_{\text{emp}}(\omega)=\frac{1}{I'}\sum_{i=1}^{I'}Q(\mathbf{z}'_i,\omega), \tag{10}$$

$$R''_{\text{emp}}(\omega)=\frac{1}{I''}\sum_{i=1}^{I''}Q(\mathbf{z}''_i,\omega) \tag{11}$$

represent respectively the training error and the testing error, calculated for some function ω. We shall also call these errors *empirical training and testing risks*.

When the context of discussion is constrained to a single fold, by $\omega_{I'}$ we define the function that minimizes the empirical training risk:

$$R'_{\text{emp}}(\omega_{I'})=\inf_{\omega\in\Omega}R'_{\text{emp}}(\omega). \tag{12}$$

[5]For the *classification* learning task, L is defined as an indicator function:

$$L\Big(f(\mathbf{x},\omega),y\Big)=\begin{cases}0, & \text{for } y=f(\mathbf{x},\omega),\\ 1, & \text{for } y\neq f(\mathbf{x},\omega),\end{cases} \tag{3}$$

and then $R(\omega)=\int_{\mathbf{x}\in\mathbf{X}}\sum_{y\in Y}L\big(f(\mathbf{x},\omega),y\big)p(\mathbf{x})P(y|\mathbf{x})\,\mathrm{d}\mathbf{x}$.

[6] In the sense of all learning tasks.

[7]Q is identical with L in the sense of their values. They differ only in the formal definition of their domains. L acts on $f(\mathbf{x},\omega)$ and y and maps them to error values, whereas Q acts directly on $\mathbf{z}$ and ω and maps them to error values.

When we need to broaden the context onto all folds, $j = 1, 2, \ldots, n$, we shall write $\omega_{I',j}$ to denote the function that minimizes the empirical training risk in the j-th fold. Therefore, the final cross-validation result—an estimate of the generalization error—is the mean of empirical testing risks R''_{emp} using functions $\omega_{I',j}$:

$$C = \frac{1}{n}\sum_{j=1}^{n} R''_{\text{emp}}(\omega_{I',j}). \tag{13}$$

1.3. Notation related to iterating over the structure. By a *structure*, a sequence of nested subsets

$$S_1 \subset S_2 \subset \cdots \subset S_K$$

is meant, where for each position $k \in \{1, 2, \ldots, K\}$ we have

$$S_k = \big\{Q(\mathbf{z}, \omega_k)\big\}_{\omega_k \in \Omega_k}, \quad 0 \le Q(\mathbf{z}, \omega_k) \le B_k$$

(a set of real-valued bounded error functions).

When the context of discussion is constrained to a single position k in the structure, we will stick to shorter notation for particular notions/objects such as, e.g., ω, ω_I, C, B, whereas when we need to broaden the context onto all positions $k \in \{1, 2, \ldots, K\}$, we shall write respectively ω_k, $\omega_{k,I}$, C_k, B_k to denote objects that come from the k-th position.

When the context of discussion requires to take into account both cross-validation and the position in the structure, we will write in particular $\omega_{k,I',j}$ to denote the function that comes from the set S_k, minimizes the empirical risk on a training set of size I', and this happens in the j-th fold of cross-validation.

1.4. Other notation details. In the paper we shall use the '$\sim$' sign with two possible meanings: (1) to denote the fact that a random variable has a certain probability distribution, e.g., $X \sim N(\mu, \sigma)$ should be read as "X is a random variable drawn from the normal distribution with mean μ and standard deviation σ"; (2) to indicate that a random variable is *similar to* or *asymptotic with* another random variable; in that case we shall skip parentheses with mean and variance, writing solely, e.g., $X \sim Y$.

In the paper we will use $N(\mu, \sigma)$ as a common notation for a normal distribution, but in other contexts we shall write N or N_k to represent the finite capacity of a set of functions for the k-th position in a structure, so a completely different notion. Recognizing the right meaning should be easy given the context and the presence or lack of parentheses after N.

2. Bounds on generalization by Vapnik

We remind some of Vapnik's results in brief.

2.1. Finite sets of functions. Let us start with the simplest case of a *finite* set with N elements being real-valued bounded functions. Vapnik (1995; 1998) shows that, with probability at least $1 - \eta$, $0 < \eta < 1$, the following bound on the true risk is satisfied:

$$\underbrace{\int_{\mathbf{Z}} Q(\mathbf{z}, \omega_I)\mathrm{d}F(\mathbf{z})}_{R(\omega_I)} \le \underbrace{\frac{1}{I}\sum_{i=1}^{I} Q(\mathbf{z}_i, \omega_I)}_{R_{\text{emp}}(\omega_I)} + B\sqrt{\frac{\ln N - \ln \eta}{2I}}. \tag{14}$$

The argument is the following:

$$\begin{aligned} &P\Big(\sup_{\omega\in\Omega} R(\omega) - R_{\text{emp}}(\omega) \ge \epsilon\Big) \\ &\quad\le \sum_{\omega\in\Omega} P\Big(R(\omega) - R_{\text{emp}}(\omega) \ge \epsilon\Big) \\ &\quad\le N \cdot \exp\Big(-\frac{2\epsilon^2 I}{B^2}\Big). \end{aligned} \tag{15}$$

The last inequality is true, since for each term in the sum, the Hoeffding inequality is satisfied. By substituting the right-hand-side by a small probability η and solving for ϵ, one obtains the bound

$$R(\omega) - R_{\text{emp}}(\omega) \le B\sqrt{\frac{\ln N - \ln \eta}{2I}},$$

which holds true with probability at least $1 - \eta$ simultaneously *for all* functions in the set, since it holds for the worst case. Hence, in particular, it holds true for the function ω_I and one gets the bound (14).

For the theorems to follow, we will denote by V the right-hand side in the Vapnik bound:

$$V = R_{\text{emp}}(\omega_I) + B\sqrt{\frac{\ln N - \ln \eta}{2I}}. \tag{16}$$

We remark that, for regression estimation, the bound (14) can be in practice tightened by using an estimate $\widehat{B}$ in place of the most pessimistic B. $\widehat{B}$ can be found, e.g., by performing just one fold of cross-validation, instead of n folds, and bounding it by the mean error on the testing set plus a square root implied by the Hoeffding inequality:

$$\widehat{B} \le R''_{\text{emp}}(\omega'_I) + B\sqrt{\frac{-\ln \eta_B}{2I''}}, \tag{17}$$

where η_B is an imposed small probability that (17) is not true. The reasoning behind this remark is that in practice typical learning algorithms, in the process of ERM, rarely produce functions $f(\mathbf{x}, \omega_I)$ having maximal possible errors within the given set of functions. Therefore, we can

insert the right-hand side of (17) into (14) in place of B and tighten the bound. If this is done, however, the probabilities for inequalities must be adjusted and become $1-\eta-\eta_B$, rather than $1-\eta$.[8]

2.2. Infinite sets of functions. The simplest case with a finite number of functions in the set was generalized by Vapnik (1995; 1998) onto *infinite* sets with a continuum of elements by introducing several notions of *capacity* for the set of functions: *entropy*, *annealed entropy*, *growth function*, *Vapnik–Chervonenkis dimension*.

Simply speaking, one should think what replacement of $\ln N$ can be made in the bound when making extension onto infinite sets. It is good to look at an infinite set of functions as an equivalent to a certain finite set of functions, in such a sense that from a continuum of functions we pick only a finite number of functions which *matter*, i.e., cause a relevant change in the risk.

First of all, it is convenient to start from the *classification* task and therefore sets of indicator error functions $Q(\mathbf{z},\omega) \in \{0,1\}$. Vapnik defines $N^\Omega(\mathbf{z}_1,\dots,\mathbf{z}_I)$ as the number of all possible *dichotomies*[9] that can be achieved on a *fixed* sample $\{\mathbf{z}_1,\dots,\mathbf{z}_I\}$ using functions from $\{Q(\mathbf{z},\omega)\}_{\omega\in\Omega}$. Obviously, $N^\Omega(\mathbf{z}_1,\dots,\mathbf{z}_I) \le 2^I$. Then, if we relax (unfix) the sample but it remains of size I and drawn from $p(\mathbf{z})$, we can think, for example, of the expected value of $\ln N^\Omega$. Vapnik introduces the following notions of *capacity*:

1. expected value of $\ln N^\Omega$—*Vapnik-Chervonenkis entropy*:

$$H^\Omega(I) = \int_{\mathbf{z}_1\in\mathbf{Z}} \cdots \int_{\mathbf{z}_I\in\mathbf{Z}} \ln N^\Omega(\mathbf{z}_1,\dots,\mathbf{z}_I)\, p(\mathbf{z}_1)\cdots p(\mathbf{z}_I)\,\mathrm{d}\mathbf{z}_1\cdots\mathrm{d}\mathbf{z}_I;$$

2. $\ln$ of expected value of N^Ω—*annealed entropy*:

$$H^\Omega_{\text{ann}}(I) = \ln \int_{\mathbf{z}_1\in\mathbf{Z}} \cdots \int_{\mathbf{z}_I\in\mathbf{Z}} N^\Omega(\mathbf{z}_1,\dots,\mathbf{z}_I)\, p(\mathbf{z}_1)\cdots p(\mathbf{z}_I)\,\mathrm{d}\mathbf{z}_1\cdots\mathrm{d}\mathbf{z}_I;$$

3. $\ln$ of supremum of N^Ω—*growth function*:

$$G^\Omega(I) = \ln \sup_{\mathbf{z}_1,\dots,\mathbf{z}_I} N^\Omega(\mathbf{z}_1,\dots,\mathbf{z}_I).$$

Vapnik (1998) proved that

$$G^\Omega(I) = \begin{cases} = \ln 2^I & \text{for } I \le h, \\ \le \ln \sum_{k=0}^h \binom{I}{k} & \text{for } I > h, \end{cases} \tag{18}$$

where h is the *Vapnik–Chervonenkis dimension*.

The VC dimension as the notion of capacity is practically useful because it is distribution-free—it does not depend on the unknown $p(\mathbf{z})$. Furthermore, Vapnik (1998) showed that

$$\begin{aligned} H^\Omega(I) &\overset{\text{(Jensen)}}{\le} H^\Omega_{\text{ann}}(I) \le G^\Omega(I) \\ &\le \ln \sum_{k=0}^h \binom{I}{k} \le \ln\Big(\frac{eI}{h}\Big)^h \\ &= h(1+\ln\frac{I}{h}). \end{aligned} \tag{19}$$

Hence the right-hand side of (19) can be suitably inserted in the bounds to replace $\ln N$.

We mention that the remaining part of generalization from infinite sets of indicator functions (classification) onto infinite sets of real-valued functions (regression estimation) can be found in the work of Vapnik (1998) and is based on the notions of a *minimal finite ϵ-net*, a *set of classifiers* for a fixed real-valued f and *a complete set of classifiers* for Ω. Still, the notion of the Vapnik-Chervonenkis dimension remains essentially the same.

It is also worth mentioning that the concept of the *minimal finite ϵ-net* is equivalent to the concept of the *ϵ-covering number*, which was studied by Bartlett *et al.* (1997) and Zhang (2002).

3. Scenario I: Cross-validation and true risks

In this section we consider the following scenario: We iterate over the structure $S_1 \subset S_2 \subset \cdots \subset S_K$ and for each its subset S_k we perform n-fold non-stratified cross-validation. We obtain a result C_k. We remind that C_k gives us an estimate of the *mean of unknown true risks* of n functions chosen by ERM in particular folds, using in each a training set of size $\frac{n-1}{n}I$:

$$C_k = \frac{1}{n}\sum_{j=1}^n R(\omega_{k,I',j}). \tag{20}$$

All those n functions can be *distinct*, but sometimes they can be repeated. This depends on whether we work respectively with an *infinite* or a *finite* set of functions, and also on the random split into training and testing subsets (remember that the cross-validation is non-stratified[10]).

[8] When joining probabilistic inequalities holding true with $1-\eta$ each, the minimum probability with which both hold true simultaneously must be $1-2\eta$, rather than $(1-\eta)^2$ (probabilistic independence case) due to possible correlations between them. It can be also viewed as a consequence of Bernoulli's inequality.

[9] For a fixed ω^*, the sequence $\big(Q(\mathbf{z}_1,\omega^*),\dots,Q(\mathbf{z}_I,\omega^*)\big)$ is a binary sequence representing correct and incorrect classifications on the given sample. With ω unfixed, i.e. going over all the possibilities from Ω, we obtain different sequences $\big(Q(\mathbf{z}_1,\omega),Q(\mathbf{z}_2,\omega),\dots,Q(\mathbf{z}_I,\omega)\big)$. The number of *distinct* sequences of this type is $N^\Omega(\mathbf{z}_1,\dots,\mathbf{z}_I)$.

[10] If the data points are distinct, the probability that exactly the same two training sets occur in two folds is $1/\binom{I}{I'}$. But after n folds, we can expect the number of non-distinct pairs of training sets to be $\binom{n}{2}/\binom{I}{I'}$, which can be a significant number.

However, this should not depend on the algorithm of the learning machine, since, to satisfy the definition of ERM, the algorithm should always provide us with the best function which minimizes the empirical risk (error on the training set).

When the procedure is finished for the whole structure, we have a sequence of results

$$C_1, C_2, \ldots, C_K,$$

and an *indication* that the optimal complexity is at the point k^*, such that

$$C_{k^*} = \min_{k\in\{1,\ldots,K\}} C_k. \tag{21}$$

Now we can use the whole data set of size I, not just $\frac{n-1}{n}I$ as in folds, and finally once again apply the ERM principle to choose the best function $f(\mathbf{x}, \omega_{k^*,I})$ as our final model.

We pose the following two important questions:

1. *What is the probability that the point k^*, indicated via cross-validation, is truly the minimum point of all unknown true risks $R(\omega_{k,I})$?*
2. *With what probability does the true minimum of all unknown true risks $R(\omega_{k,I})$ fall into the neighbourhood of point k^*, indicated via cross-validation, with a side Δ?*

In other words, we want to know something about the credibility of our result k^* as being supposedly the point of optimal complexity, or at least we want to know how much we could have missed about it.

We define the notion of neighbourhood for our purposes.

Definition 1. The *neighbourhood* U of point k^* with a side Δ is

$$U(k^*, \Delta) = \{k\colon |k - k^*| \le \Delta\}. \tag{22}$$

The complement of the neighbourhood is

$$\bar{U}(k^*, \Delta) = \{k\colon |k - k^*| > \Delta\}. \tag{23}$$

We now state two theorems which answer the posed questions in such a way that they give minimal (pessimistic) values of the probabilities wanted.

Theorem 1. *Let $S_1 \subset S_2 \subset \cdots \subset S_K$ be a structure of nested sets of real-valued bounded functions:*

$$S_k = \big\{Q(\mathbf{z}, \omega_k)\big\}_{\omega_k\in\Omega_k}, \quad 0 \le Q(\mathbf{z}, \omega_k) \le B_k.$$

Let each element S_k of the structure have a finite capacity N_k, i.e., a finite number of functions in the case of finite sets in the structure or a finite Vapnik–Chervonenkis dimension in the case of infinite sets. Let $C_1, C_2, \ldots, C_K$ be a sequence of results from an n-fold non-stratified cross-validation procedure performed for this structure. Suppose the minimum of cross-validation result is reached at the point k^*:

$$C_{k^*} = \min_{k\in\{1,\ldots,K\}} C_k.$$

Then the minimal probability that the point k^*, indicated via cross-validation, is truly the minimum point of unknown true risks $R(\omega_{k,I})$ and can be calculated as follows:

$$P\Big(R(\omega_{k^*,I}) = \min_{k\in\{1,\ldots,K\}} R(\omega_{k,I})\Big) = \int_{-\infty}^{\infty}\Big(\prod_{\substack{k\in\{1,\ldots,K\}\\ k\ne k^*}}\int_{r_{k^*}}^{\infty} p_k(r_k)\,\mathrm{d}r_k\Big)p_{k^*}(r_{k^*})\,\mathrm{d}r_{k^*}, \tag{24}$$

where p_k are normal probability densities:

$$p_k(r) = \frac{1}{\frac{1}{\sqrt{n}}\sqrt{{\sigma_{k1}}^2 + {\sigma_{k2}}^2}\sqrt{2\pi}}\exp\left(-\frac{(r-C_k)^2}{\frac{2}{n}({\sigma_{k1}}^2 + {\sigma_{k2}}^2)}\right) \tag{25}$$

with the constants

$$\begin{aligned}\sigma_{k1} &= \frac{B_k\sqrt{n}}{a_{1-\frac{\eta}{2}}}\sqrt{\frac{-\ln\frac{\eta}{2}}{2I}},\\ \sigma_{k2} &= \frac{B_k}{a_{1-\frac{\eta}{2}}}\Bigg(\sqrt{\frac{n}{n-1}}\sqrt{\frac{-\ln\frac{\eta}{6}}{2I}}\\ &\quad + \Big(\sqrt{\frac{n}{n-1}}+1\Big)\sqrt{\frac{\ln N_k - \ln\frac{\eta}{6}}{2I}}\Bigg).\end{aligned} \tag{26}$$

$a_{1-\frac{\eta}{2}}$ denotes a quantile of order $1-\frac{\eta}{2}$ from $N(0,1)$ for any small $\eta > 0$. Normal distributions are approximations of unknown true risks distributions with the uniform[11] error of order $O\left(\left(1+\frac{1}{\sqrt{n-1}}+\sqrt{n}\right)\frac{1}{\sqrt{I}}\right)$.

In the next theorem we state only the thesis, as the assumptions are the same as in Theorem 1.

Theorem 2. *The minimal probability that the true minimum of unknown true risks $R(\omega_{k,I})$ falls into the neighbourhood $U(k^*, \Delta)$ of the point k^*, indicated via cross-validation, can be calculated as follows:*

$$P\Big(\arg\min_{k\in\{1,\ldots,K\}} R(\omega_{k,I}) \in U(k^*, \Delta)\Big) = \sum_{k\in U(k^*,\Delta)}\int_{-\infty}^{\infty}\Big(\prod_{\substack{l\in\{1,\ldots,K\}\\ l\ne k}}\int_{r_k}^{\infty} p_l(r_l)\,\mathrm{d}r_l\Big) \cdot p_k(r_k)\,\mathrm{d}r_k, \tag{27}$$

[11] In the sense of the supremum of errors for the distribution cumulative function taken over all r. Details are given in Appendix B.

where p_l, p_k are normal probability densities defined as in (25) *with the uniform error of order* $O\left(\left(1+\frac{1}{\sqrt{n-1}}+\sqrt{n}\right)\frac{1}{\sqrt{I}}\right)$.

In theorems, the inner expression under the integral of type $\int_{r_k}^{\infty} p_l(r_l)\,\mathrm{d}r_l$ could also be written down, for example, as $P(r_k < R(\omega_{l,I}))$, denoting the probability that the value of $R(\omega_{l,I})$ is greater than a treshold r_k—the outer integral variable.

The proof of Theorems 1 and 2 will be carried out firstly by proving two lemmas which justify the form of densities p_k, and secondly by showing the right technique to calculate the final probabilities on the basis of these densities. These two parts will conclude the proof.

In the lemmas (and corollaries), we apply the central limit theorem in several places and we approximate a certain unknown distribution by a normal distribution. With respect to the sample size I and the number of cross-validation folds n, the order of the approximation uniform error is $O\left(\left(1+\frac{1}{\sqrt{n-1}}+\sqrt{n}\right)\frac{1}{\sqrt{I}}\right)$. We give more details about the accuracy of CLT normal approximations in Appendix B on the basis of the Berry–Esséen theorem.

Additionally we shall say that the approximation is *pessimistic*. First of all, this means that both distributions are close to each other in the sense of some metric (i.e., their density functions and cumulative densities are close), but more importantly this means that the approximating normal distribution is of greater uncertainty than the approximated distribution. More formally, given $0<\eta<1$ and two close distributions A_*, A with densities p_{A_*}, p_A, we shall say that A_* is *pessimistically* approximated by A if and only if for all quantiles $a_{1-\frac{\eta_0}{2}}$ where $\eta_0 \leq \eta$, taken from A, the condition

$$\int_{-a_{1-\frac{\eta_0}{2}}}^{a_{1-\frac{\eta_0}{2}}} p_{A_*}(x)\,\mathrm{d}x \geq \int_{-a_{1-\frac{\eta_0}{2}}}^{a_{1-\frac{\eta_0}{2}}} p_A(x)\,\mathrm{d}x \tag{28}$$

is satisfied.

The notion 'minimal probability' used in both theorems is justified by Theorem 4, given in Appendix A, where we prove that by tightening variances for *any* position in the structure the probabilities (24) and (27) can only by improved, not worsened, which might not be intuitively obvious.

Lemma 1. *For any $\eta > 0$, arbitrarily small, the distribution of $R''_{emp}(\omega_{I'})$ in each single fold can be pessimistically approximated by the normal distribution with the following expected value and standard deviation:*

$$R''_{emp}(\omega_{I'}) \sim N\left(R(\omega_{I'}), \frac{B\sqrt{n}}{a_{1-\frac{\eta}{2}}}\sqrt{\frac{-\ln\frac{\eta}{2}}{2I}}\right), \tag{29}$$

where $a_{1-\frac{\eta}{2}}$ is a quantile of order $1-\frac{\eta}{2}$ from $N(0,1)$.

Proof. For a fixed function $f(\mathbf{x},\omega_{I'})$ chosen in a single fold via ERM, the error value $Q(\mathbf{z},\omega_{I'})$ for any testing sample point $\mathbf{z}=(\mathbf{x},y)$, taken at random from the distribution with the joint density $p(\mathbf{z})$, has a certain probability distribution around the value of true risk $R(\omega_{I'})$ (expected value) with a certain unknown variance σ. Since $R''_{\text{emp}}(\omega_{I'})$ arises as a mean, thus also a sum, of I'' *independent* results, then by means of the *central limit theorem* we can approximate it by a normal distribution with a standard deviation equal to

$$\frac{1}{I''}\sqrt{\sum_{i=1}^{I''}\sigma^2} = \frac{\sigma}{\sqrt{I''}}.$$

Hence

$$R''_{\text{emp}}(\omega_{I'}) \sim N\left(R(\omega_{I'}), \frac{\sigma}{\sqrt{I''}}\right). \tag{30}$$

The pessimistic σ can be derived by using the *Hoeffding inequality* and joining it with an appropriate equality implied by the normal distribution. We write respectively

$$P\left(|R(\omega_{I'}) - R''_{\text{emp}}(\omega_{I'})| \leq B\sqrt{\frac{-\ln\frac{\eta}{2}}{2I''}}\right) \geq 1-\eta, \tag{31}$$

$$P\left(|R(\omega_{I'}) - R''_{\text{emp}}(\omega_{I'})| \leq a_{1-\frac{\eta}{2}}\frac{\sigma}{\sqrt{I''}}\right) = 1-\eta. \tag{32}$$

By comparison, we see that the condition for σ is

$$\sigma \leq \frac{B}{a_{1-\frac{\eta}{2}}}\sqrt{\frac{-\ln\frac{\eta}{2}}{2}}, \tag{33}$$

so it is sufficient to pessimistically set up σ to the right-hand-side of (33), in the sense that for this value the probability measure of the unknown distribution of $R''_{\text{emp}}(\omega_{I'})$ contained up to the given quantile is the same or greater than the probability measure in the known normal distribution.

Finally, by inserting $I''=\frac{1}{n}I$, we have that with probability at least $1-\eta$

$$R''_{\text{emp}}(\omega_{I'}) \sim N\left(R(\omega_{I'}), \frac{B\sqrt{n}}{a_{1-\frac{\eta}{2}}}\sqrt{\frac{-\ln\frac{\eta}{2}}{2I}}\right). \tag{34}$$

■

Now we state a lemma which shows a probabilistic relationship between true risks: $R(\omega_{I'})$ from any single fold and $R(\omega_I)$ (when using the whole data set).

Lemma 2. *For any $\eta > 0$, arbitrarily small, with probability $1-6\eta$ or greater, the following two inequalities,*

bounding $R(\omega_{I'})$ for any fold, simultaneously hold true:

$$\begin{aligned} & R(\omega_I) - B\sqrt{\frac{\ln N - \ln\eta}{2I}} \\ & \le R(\omega_{I'}) \\ & \le R(\omega_I) + B\sqrt{\frac{n}{n-1}}\sqrt{\frac{-\ln\eta}{2I}} \\ & \quad + B\Big(\sqrt{\frac{n}{n-1}}+1\Big)\sqrt{\frac{\ln N - \ln\eta}{2I}}, \end{aligned} \tag{35}$$

where N stands for a suitable notion of capacity for the given set of functions $\{Q(\mathbf{z},\omega)\}_{\omega\in\Omega}$.

Proof. The following four bounds are true with probability at least $1-\eta$ each:

$$R(\omega_I) \le R_{\text{emp}}(\omega_I) + B\sqrt{\frac{\ln N - \ln\eta}{2I}}, \tag{36}$$

$$R_{\text{emp}}(\omega_I) \le R(\omega_I) + B\sqrt{\frac{\ln N - \ln\eta}{2I}}, \tag{37}$$

$$R(\omega_{I'}) \le R'_{\text{emp}}(\omega_{I'}) + B\sqrt{\frac{\ln N - \ln\eta}{2I'}}, \tag{38}$$

$$R'_{\text{emp}}(\omega_{I'}) \le R(\omega_{I'}) + B\sqrt{\frac{\ln N - \ln\eta}{2I'}}. \tag{39}$$

The first two are one-side versions of the Vapnik bound on true risk, see (14), when using the whole data set of size I, while the second two are analogical when using a smaller training set of size $I' = \frac{n-1}{n}I$ in a single fold.

We write the following sequence of inequalities:

$$R'_{\text{emp}}(\omega_{I'}) \le R'_{\text{emp}}(\omega_I) \le R_{\text{emp}}(\omega_I) + B\sqrt{\frac{-\ln\eta}{2I'}}. \tag{40}$$

The first one is true with probability 1 by the definition of $\omega_{I'}$, the second one is a Hoeffding inequality, true with probability at least $1-\eta$ for the fixed function ω_I.

By joining (38) and (40), we obtain with probability at least $1-2\eta$

$$R(\omega_{I'}) \le R_{\text{emp}}(\omega_I) + B\sqrt{\frac{-\ln\eta}{2I'}} + B\sqrt{\frac{\ln N - \ln\eta}{2I'}}. \tag{41}$$

By joining this further with (37) and plugging $I' = \frac{n-1}{n}I$, we obtain with probability at least $1-3\eta$

$$\begin{aligned} R(\omega_{I'}) & \le R(\omega_I) + B\sqrt{\frac{n}{n-1}}\sqrt{\frac{-\ln\eta}{2I}} \\ & \quad + B\Big(\sqrt{\frac{n}{n-1}}+1\Big)\sqrt{\frac{\ln N - \ln\eta}{2I}}. \end{aligned} \tag{42}$$

This proves the right hand side bound in the lemma.

To prove the left-hand side, we write the following sequence of inequalities:

$$\begin{aligned} R'_{\text{emp}}(\omega_{I'}) & \ge R_{\text{emp}}(\omega_{I'}) + B\sqrt{\frac{\ln N - \ln\eta}{2I'}} \\ & \ge R_{\text{emp}}(\omega_I) + B\sqrt{\frac{\ln N - \ln\eta}{2I'}}. \end{aligned} \tag{43}$$

The first one is a bound similar to Vapnik's[12] and it is true with probability at least $1-\eta$, while the second is true with probability 1 from the definition of ω_I.

By joining (36) and (43), we obtain with probability at least $1-2\eta$

$$\begin{aligned} R(\omega_I) & \le R'_{\text{emp}}(\omega_{I'}) - B\sqrt{\frac{\ln N - \ln\eta}{2I'}} \\ & \quad + B\sqrt{\frac{\ln N - \ln\eta}{2I}}. \end{aligned} \tag{44}$$

By joining this further with (39) and plugging $I' = \frac{n-1}{n}I$, we obtain with probability at least $1-3\eta$

$$\begin{aligned} R(\omega_I) & \le R(\omega_{I'}) + B\sqrt{\frac{n}{n-1}}\sqrt{\frac{\ln N - \ln\eta}{2I}} \\ & \quad - B\sqrt{\frac{n}{n-1}}\sqrt{\frac{\ln N - \ln\eta}{2I}} \\ & \quad + B\sqrt{\frac{\ln N - \ln\eta}{2I}}. \end{aligned} \tag{45}$$

As we see the first two summands cancel out and this proves the left-hand side bound in the lemma. ■

Owing to Lemma 2, we can pessimistically approximate the distribution of $R(\omega_{I'})$ by a normal distribution with the expected value $R(\omega_I)$, which is a constant, and a standard deviation determined by the right-hand side of the lemma, since it is broader than the left-hand side. We remind the right-hand side is true with probability at least $1-3\eta$, but for further deliberations we need to put the probabilities (and quantiles) in agreement to the level $1-\eta$, so we pay attention to doing so. Following the lemma, we write the probabilistic inequality

$$\begin{aligned} & P\bigg(|R(\omega_{I'}) - R(\omega_I)| \le B\sqrt{\frac{n}{n-1}}\sqrt{\frac{-\ln\frac{\eta}{6}}{2I}} \\ & \quad + B\Big(\sqrt{\frac{n}{n-1}}+1\Big)\sqrt{\frac{\ln N - \ln\frac{\eta}{6}}{2I}}\bigg) \ge 1-\eta, \end{aligned} \tag{46}$$

[12] The measure R_{emp} corresponds by analogy to the measure R in the original Vapnik bound, and the measure R'_{emp} corresponds by analogy to R_{emp} therein. Obviously, R is defined on an infinite and continuous space $\mathbf{Z} = \mathbf{X} \times Y$, whereas R_{emp} is defined on a discrete and finite sample $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_I\}$, but still from the perspective of a single cross-validation fold we may view $R_{\text{emp}}(\omega_I)$ as the "target" minimal *error expectation* and $R'_{\text{emp}}(\omega_{I'})$ as the observed relative *mean error*—an estimate of the expectation.

and we compare it with a suitable equality implied by the normal distribution

$$P\Big(|R(\omega_{I'}) - R(\omega_I)| \le a_{1-\frac{\eta}{2}}\sigma_2\Big) = 1-\eta. \tag{47}$$

We name the standard deviation σ_2 for a purpose. We see that, pessimistically, σ_2 must be at least

$$\sigma_2 = \frac{1}{a_{1-\frac{\eta}{2}}}\Bigg(B\sqrt{\frac{n}{n-1}}\sqrt{\frac{-\ln\frac{\eta}{6}}{2I}} + B\Big(\sqrt{\frac{n}{n-1}}+1\Big)\sqrt{\frac{\ln N - \ln\frac{\eta}{6}}{2I}}\Bigg). \tag{48}$$

Corollary 1. *For any $\eta > 0$, arbitrarily small, we can pessimistically approximate $R(\omega_{I'})$ in each fold by the following normal distribution:*

$$R(\omega_{I'}) \sim N\Big(R(\omega_I), \sigma_2\Big). \tag{49}$$

Let us look back at the bottom line of Lemma 1. We have that

$$R''_{\text{emp}}(\omega_{I'}) \sim N\Bigg(R(\omega_{I'}), \underbrace{\frac{B\sqrt{n}}{a_{1-\frac{\eta}{2}}}\sqrt{\frac{-\ln\frac{\eta}{2}}{2I}}}_{\sigma_1}\Bigg), \tag{50}$$

whereas from Corollary 1 we have that

$$R(\omega_{I'}) \sim N\Big(R(\omega_I), \sigma_2\Big).$$

We see that $R''_{\text{emp}}(\omega_{I'}) \sim R(\omega_{I'}) \sim R(\omega_I)$, meaning that in a single fold of cross-validation the *empirical testing risk* calculated for a function $\omega_{I'}$ *is similar to* the unknown *true risk* for this function, i.e., estimates it with a certain deviation, and in turn this true risk *is similar to* the *true risk* of ω_I, i.e., the function that we would choose by ERM if the whole data set was taken into account, not just the training set of the fold. This can be regarded as the nesting of random variables, and we can write

$$R''_{\text{emp}}(\omega_{I'}) \sim N\Bigg(R(\omega_I), \sqrt{{\sigma_1}^2 + {\sigma_2}^2}\Bigg). \tag{51}$$

The fact that variances should be summed up for nested random variables is demonstrated in Appendix C. By taking the mean after n independent folds of cross-validation, again by means of CLT, we write the final consequence which gives us a distribution with a standard deviation smaller by factor $1/\sqrt{n}$.

Corollary 2. *For any $\eta > 0$, arbitrarily small, the final result of cross-validation for the k-th position in the structure can be approximated by the normal distribution with the following expected value and standard deviation:*

$$C_k \sim N\Bigg(R(\omega_{k,I}), \frac{1}{\sqrt{n}}\sqrt{{\sigma_{k1}}^2 + {\sigma_{k2}}^2}\Bigg), \tag{52}$$

where values σ_{k1}, σ_{k2} are defined for the k-th position in the structure according to the formulas (50) *and* (48).

For a given experiment, we do know in fact the realizations of each C_k, i.e., we know their exact values, since we have them measured, whereas unknown are the *true risks* $R(\omega_{k,I})$. Nevertheless, by symmetry we can probabilistically assess the value of $R(\omega_{k,I})$ knowing a C_k, for any desired probability $1-\alpha$:

$$P\Bigg(|C_k - R(\omega_{k,I})| \le a_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n}}\sqrt{{\sigma_{k1}}^2 + {\sigma_{k2}}^2}\Bigg) \ge 1-\alpha. \tag{53}$$

Therefore, although each $R(\omega_{k,I})$ is in fact a constant, we can regard it as a random variable with respect to a C_k, i.e.,

$$R_k(\omega_I) \sim N\Big(C_k(\omega_I), \frac{1}{\sqrt{n}}\sqrt{{\sigma_{k1}}^2 + {\sigma_{k2}}^2}\Big).$$

This fact, in conjunction with the technique to calculate probabilities (by suitable integrals shown in the next section), implies proving Theorems 1 and 2.

4. Calculation of probabilities

Let $p(r_1, r_2, \ldots, r_K)$ be the K-dimensional density function. It represents the joint probability distribution of the values of true risks $R(\omega_{k,I})$ for the whole structure, i.e., taking into account all positions $k \in \{1, 2, \ldots, K\}$. Owing to independence, the joint density is the product of one-dimensional densities:

$$p(r_1, r_2, \ldots, r_K) = p_1(r_1)p_2(r_2)\cdots p_K(r_K), \tag{54}$$

which are normal densities with expectations and standard deviations defined by Theorems 1, 2, see the exemple in Fig. 2. To calculate the probabilities wanted, we need to suitably integrate the joint density $p(r_1, r_2, \ldots, r_K)$. It is convenient to demonstrate the right technique using a convention called the *Iverson notation* (Knuth, 1997; Graham *et al.*, 2002):

$$[s] = \begin{cases} 1 & \text{when } s \text{ is true,} \\ 0 & \text{when } s \text{ is false,} \end{cases}$$

where s is an arbitrary statement. In other words, we shall integrate $p(r_1, r_2, \ldots, r_K)$ over the space of values of all true risks and we shall selectively turn on and off suitable subsets of this space with $\{0, 1\}$ statements.

The following formula answers the first question that we posed in the former section, namely

1. *What is the probability that the point k^*, indicated via cross-validation, is truly the minimum point of all unknown true risks $R(\omega_{k,I})$?*

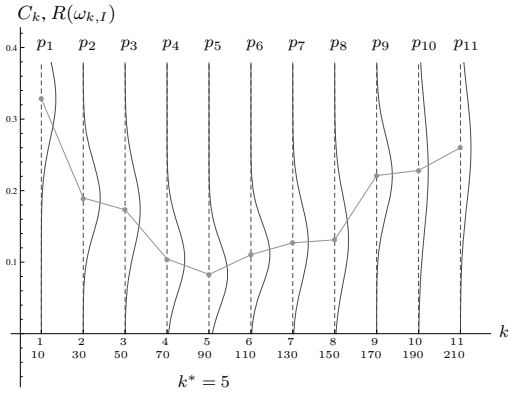

Fig. 2. Example result of complexity selection procedure via cross-validation. Optimal complexity suggested at point $k^* = 5$. Probability distribution densities of $R(\omega_{k,I})$ drawn symbolically along vertical axes.

$$
\begin{aligned}
&P\Big(R(\omega_{k^*,I}) = \min_{k\in\{1,\ldots,K\}} R(\omega_{k,I})\Big) \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}[r_{k^*} < r_1]\cdots[r_{k^*} < r_{k^*-1}] \\
&\qquad\qquad \times [r_{k^*} < r_{k^*+1}]\cdots[r_{k^*} < r_K] \\
&\qquad\qquad \times p(r_1,\ldots,r_K)\,\mathrm{d}r_1\cdots\mathrm{d}r_K \\
&= \int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty}[r_{k^*} < r_1]p_1(r_1)\,\mathrm{d}r_1\cdots \\
&\qquad\qquad \times \int_{-\infty}^{\infty}[r_{k^*} < r_K]p_K(r_K)dr_K\Big)p_{k^*}(r_{k^*})\,\mathrm{d}r_{k^*} \\
&= \int_{-\infty}^{\infty}\Big(\int_{r_{k^*}}^{\infty}p_1(r_1)\,\mathrm{d}r_1\cdots \\
&\qquad\qquad \times \int_{r_{k^*}}^{\infty}p_K(r_K)\,\mathrm{d}r_K\Big)p_{k^*}(r_{k^*})\,\mathrm{d}r_{k^*} \\
&= \int_{-\infty}^{\infty}\Big(\prod_{\substack{k\in\{1,\ldots,K\}\\ k\neq k^*}} P\Big(r_{k^*} < R(\omega_{k,I})\Big)\Big)p_{k^*}(r_{k^*})\,\mathrm{d}r_{k^*}.
\end{aligned}
\tag{55}
$$

It is worth stating that, obviously, the values of true risk cannot be negative, so it seems wrong to take integrals from $-\infty$. However, since we agreed to approximate the joint distribution by a normal, we have to follow it formally and integrate from $-\infty$, so that the probabilities can sum up to 1 (if calculated for all positions possible to be k^*).

Now, we want to answer the second important question posed in the former section, namely,

2. *With what probability does the true minimum of all unknown true risks $R(\omega_{k,I})$ fall into the neighbourhood of point k^*, indicated via cross-validation, with a side Δ?*

To answer it, let us begin with a small case example. Imagine that in the process of cross-validation with $k \in \{1,\ldots,7\}$ we obtained

$$C_1 > C_2 > C_3 > C_4 < C_5 < C_6 < C_7,$$

and hence $k^* = 4$. Say, we impose the neighbourhood with $\Delta = 1$, hence $U(k^*,\Delta) = \{3,4,5\}$, $\bar{U}(k^*,\Delta) = \{1,2,6,7\}$. To calculate the desired probability, we need to use the Bayesian total probability formula, with three conditional events, disjoint pair-wise, accounting for cases where each $R(\omega_{k,I})$ in the $U(k^*,\Delta)$ is the smallest, i.e., $R(\omega_{3,I}) < R(\omega_{4,I}), R(\omega_{5,I})$ or $R(\omega_{4,I}) < R(\omega_{3,I}), R(\omega_{5,I})$ or $R(\omega_{5,I}) < R(\omega_{3,I}), R(\omega_{4,I})$. Again, we write a suitable integral of the joint probability density with Iverson zero/one statements:

$$
\begin{aligned}
&P\Big(\arg\min_{k\in\{1,\ldots,7\}} R(\omega_{k,I}) \in U(k^*=4,\Delta=1)\Big) \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\Big([r_3 < r_4, r_5][r_3 < r_1, r_2, r_6, r_7] \\
&\qquad + [r_4 < r_3, r_5][r_4 < r_1, r_2, r_6, r_7] \\
&\qquad + [r_5 < r_3, r_4][r_5 < r_1, r_2, r_6, r_7]\Big) \\
&\qquad\qquad \times p(r_1,\ldots,r_K)\,\mathrm{d}r_1\cdots\mathrm{d}r_K \\
&= \int_{-\infty}^{\infty}\Big(\iint_{r_3}^{\infty} p(r_4,r_5)\,\mathrm{d}r_4\,\mathrm{d}r_5 \\
&\qquad \times \iiiint_{r_3}^{\infty} p(r_1,r_2,r_6,r_7)\,\mathrm{d}r_1\,\mathrm{d}r_2\,\mathrm{d}r_6\,\mathrm{d}r_7\Big)p(r_3)\,\mathrm{d}r_3 \\
&+ \int_{-\infty}^{\infty}\Big(\iint_{r_4}^{\infty} p(r_3,r_5)\,\mathrm{d}r_3\,\mathrm{d}r_5 \\
&\qquad \times \iiiint_{r_4}^{\infty} p(r_1,r_2,r_6,r_7)\,\mathrm{d}r_1\,\mathrm{d}r_2\,\mathrm{d}r_6\,\mathrm{d}r_7\Big)p(r_4)\,\mathrm{d}r_4 \\
&+ \int_{-\infty}^{\infty}\Big(\iint_{r_5}^{\infty} p(r_3,r_4)\,\mathrm{d}r_3\,\mathrm{d}r_4 \\
&\qquad \times \iiiint_{r_5}^{\infty} p(r_1,r_2,r_6,r_7)\,\mathrm{d}r_1\,\mathrm{d}r_2\,\mathrm{d}r_6\,\mathrm{d}r_7\Big)p(r_5)\,\mathrm{d}r_5.
\end{aligned}
\tag{56}
$$

From the small case example, we see that, given a neighbourhood $U(k^*,\Delta)$ and its complement $\bar{U}(k^*,\Delta)$, the general formula is

$$
\begin{aligned}
&P\Big(\arg\min_{k\in\{1,\ldots,K\}} R(\omega_{k,I}) \in U(k^*,\Delta)\Big) \\
&= \sum_{k\in U(k^*,\Delta)}\int_{-\infty}^{\infty}\Big(\prod_{\substack{j\in U(k^*,\Delta)\\ j\neq k}} P\Big(r_k < R(\omega_{j,I})\Big)
\end{aligned}
$$

$$\times \prod_{j\in\bar{U}(k^*,\Delta)} P\Big(r_k < R(\omega_{j,I})\Big)\Big) \times p_k(r_k)\,\mathrm{d}r_k. \quad (57)$$

Since $(U(k^*,\Delta)\setminus\{k\})\cup\bar{U}(k^*,\Delta)=\{1,\ldots,K\}\setminus\{k\}$, we join products under the integral and rewrite the formula as

$$P\Big(\arg\min_{k\in\{1,\ldots,K\}} R_k(\omega_I)\in U(k^*,\Delta)\Big) = \sum_{k\in U(k^*,\Delta)}\int_{-\infty}^{\infty}\Big(\prod_{\substack{j\in\{1,\ldots,K\}\\ j\neq k}} P\Big(r_k < R(\omega_{j,I})\Big)\Big) \times p_k(r_k)\,\mathrm{d}r_k. \quad (58)$$

For a moment we come back to the formula (55). It takes into account all positions k in the structure as "competitors" of k^* to be minimum. One could be interested in calculating an additional probability but for a narrower domain—only a certain neighbourhood of point k^*, say with a side Δ. In other words, one could ask: *What is the probability that the unknown true risk at point k^* is the smallest within its neighbourhood $U(k^*,\Delta)$?* The resulting formula differs from (55) only in the set of indices for which the product is calculated:

$$P\Big(R(\omega_{k^*,I}) = \min_{k\in U(k^*,\Delta)} R(\omega_{k,I})\Big) = \int_{-\infty}^{\infty}\Big(\prod_{\substack{k\in U(k^*,\Delta)\\ k\neq k^*}} P\Big(r_{k^*} < R(\omega_{k,I})\Big)\Big) p_{k^*}(r_{k^*})\,\mathrm{d}r_{k^*}. \quad (59)$$

Since the set of indices for the product is narrower, the probability (59) is greater than (55).

5. Experiments for Scenario I: Minima of cross-validation and true risks

5.1. Set of functions. The form of f functions, $f\colon[0,1]^2\to[-1,1]$, was Gaussian-like:

$$f(\mathbf{x},\overbrace{w_0,w_1,\ldots,w_M}^{\omega}) = \max\Bigg\{-1,\min\Big\{1,w_0 + \sum_{m=1}^{M} w_m\exp\Big(-\frac{\|\mathbf{x}-\mu_m\|^2}{2{\sigma_m}^2}\Big)\Big\}\Bigg\}, \quad (60)$$

where centers μ_m and widths σ_m were generated at random[13] and remained *fixed*. Therefore we have a set of functions linear in parameters $(w_0,w_1,\ldots,w_M)$. As one can see, values of f were purposely constrained by ±1. Examples of functions from this set are shown in Fig. 3.

5.2. System and data sets. As a system $y(\mathbf{x})$ we picked at random a function from a class similar to (60) but *broader*, in the sense that the number M was greater and the range of randomness on σ_m was larger. Data sets for experiments were taken by sampling the system according to the joint probability density $p(\mathbf{x},y)=p(\mathbf{x})p(y|\mathbf{x})$ where we imposed $p(\mathbf{x})=1$, i.e., a uniform distribution on the domain $[0,1]^2$, and

$$p(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_\epsilon}\exp\left(-\frac{(y-y(\mathbf{x}))^2}{2{\sigma_\epsilon}^2}\right),$$

normal noise with a standard deviation $\sigma_\epsilon=0.1$, see Fig. 4.

5.3. Algorithm of the learning machine. The learning machine was trained by using least-squares criterion. We remark that, obviously, other learning approaches can be used here, e.g., the maximum likelihood criterion, the SVM regression criterion (Vapnik, 1998; Korzeń and Klęsk, 2008). Let us denote by $g_k(\mathbf{x})$ the basis $\exp\big(-\|\mathbf{x}-\mu_m\|^2/(2{\sigma_m}^2)\big)$. If we calculate the matrix of basis values at data points

$$G = \begin{pmatrix} 1 & g_1(\mathbf{x}_1) & g_2(\mathbf{x}_1) & \cdots & g_M(\mathbf{x}_1)\\ 1 & g_1(\mathbf{x}_2) & g_2(\mathbf{x}_2) & \cdots & g_M(\mathbf{x}_2)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & g_1(\mathbf{x}_I) & g_2(\mathbf{x}_I) & \cdots & g_M(\mathbf{x}_I)\end{pmatrix}, \quad (61)$$

then we can find the optimal vector of w coefficients as follows:

$$(w_0,w_1,\ldots,w_M)^T = (G^TG)^{-1}G^TY, \quad (62)$$

where $Y=(y_1,y_2,\ldots,y_I)^T$ is the vector of training target values.

5.4. Results of experiments. In Table 1 we show results of experiments. For each experiment, 100 repetitions were carried out[14]. For each repetition a data set of given size I was drawn from the fixed distribution with $p(\mathbf{z})$ density, see Fig. 4. When an experiment was completed, we looked at *minimal probabilities* calculated according to Theorems 1 and 2 of three events:

1. *Is true minimum at point k^*?*

[13] Random intervals: $\mu_m\in[0,1]^2$, $\sigma_m\in[0.05,0.5]/\ln(m+1)$. The basis narrows down with m.

[14] Experiments were performed on a computer with: 2 GHz processor, 1 GB of RAM, using *Mathematica 6.0*. Depending on the complexity of experiment: the *structure*, the size of the sample, the number of cross-validation folds—100 of repetitions were taken from about 3 h to about 12 h of duration.

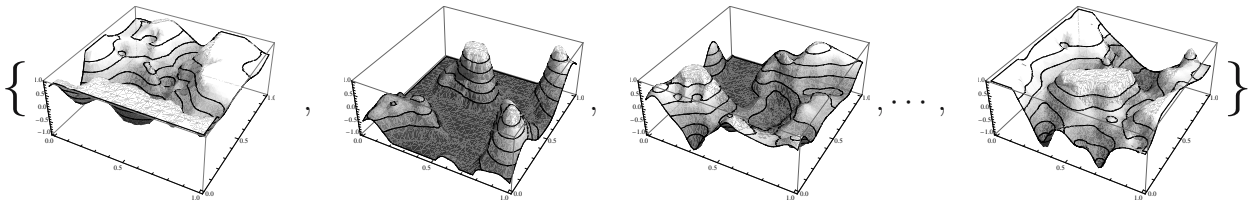

Fig. 3. Illustration of the set of approximating functions used in experiments for regression estimation.

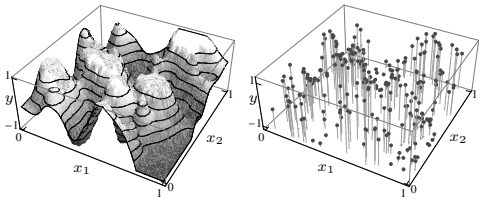

Fig. 4. System (regression function) and its data points.

2. *Is true minimum in* $U(k^*, 1)$*?*

3. *Is true minimum in* $U(k^*, 2)$*?*

and for comparison we looked at the obtained *frequencies* ν of these events observed in the 100 repetitions. Obviously, we can have frequencies of true risks only because it is a controlled experiment and we explicitly imposed $p(\mathbf{z})$, and that is why we can calculate true risks. We remind again that in practice $p(\mathbf{z})$ is *unknown* and therefore true risks are also *unknown*. From that point of view the ability to calculate probabilities stating things about true risks without knowing $p(\mathbf{z})$ is interesting.

As regards the order in which experiments are shown in the table, firstly, we change the number of cross-validation folds, secondly, we change the structure from less to more dense, and thirdly, we increase the sizes of data sets. We remark that the calculated probabilities could differ among repetitions due to slightly different data sets (noise density $p(y|\mathbf{x})$), different random splits into training and testing sets (per fold) and thus different C_k values obtained, whereas the observed frequency is just one number—a constant. Because of this fact in the table we show the means of calculated minimal probabilities over all repetitions. In Fig. 5 we show single examples of SRM plots corresponding to experiments in the table, whereas in Fig. 6 we show example models obtained for different positions in the structure.

Here are some comments on experimental results for this scenario:

1. Calculated minimal probabilities were fairly close to observed frequencies and frequencies surpassed them—empirical confirmation of the theorems.

2. Results for sparser structures obtained greater measures of probability and frequency than for denser structures when looking at a fixed Δ—it is an intuitive result. One can indicate optimal complexity more confidently in a sparser structure.

3. Apart from one exception, 20-fold cross-validation led to higher estimates of minimal probabilities than 10-fold cross-validation, but as regards observed frequencies such a property does not seem to be true.

6. Scenario II: Vapnik bounds and cross-validation

In this section we consider the following scenario: We iterate over the structure $S_1 \subset S_2 \subset \cdots \subset S_K$ and for each its subset S_k we do not perform cross-validation, we calculate only Vapnik's bound on true risk:

$$\overbrace{\int_{\mathbf{Z}} Q(\mathbf{z}, \omega_{k,I}) \mathrm{d}F(\mathbf{z})}^{R(\omega_{k,I})} \leq \overbrace{\frac{1}{I}\sum_{i=1}^{I} Q(\mathbf{z}_i, \omega_{k,I})}^{R_{\text{emp}}(\omega_{k,I})} + B_k \sqrt{\frac{\ln N_k - \ln \eta}{2I}}, \qquad (63)$$

which holds true with probability at least $1 - \eta$. We shall denote by V_k the right-hand side of the bound.

When the procedure is finished for the whole structure, we have a sequence of results

$$V_1, V_2, \ldots, V_K,$$

and an *indication* that the optimal complexity is at the point k^*, such that

$$V_{k^*} = \min_{k \in \{1,\ldots,K\}} V_k. \qquad (64)$$

Therefore, as our final model we choose the function $f(\mathbf{x}, \omega_{k^*,I})$ for which the guaranteed true risk according to (63) is minimal.

Since we do not perform n-fold cross-validation, this scenario is $O(n)$ times faster than Scenario I but less accurate. Therefore, we pose the following questions:

Table 1. Results of experiments for Scenario I. In each row there are shown the data set size, the number of cross-validation folds, the structure, the calculated minimal probabilities and observed frequencies of the events considered. For all results $\eta = 0.05$, $\eta_B = 0.25$.

no.	I	n	number of terms M (successive points in the structure)	*Is true minimum at point k^*?* min. P	*Is true minimum at point k^*?* observed ν	*Is true minimum in $U(k^*,1)$?* min. P	*Is true minimum in $U(k^*,1)$?* observed ν	*Is true minimum in $U(k^*,2)$?* min. P	*Is true minimum in $U(k^*,2)$?* observed ν
1	2	3	4	5	6	7	8	9	10
1	250	10	$\{10, 40, \ldots, 250\}$	0.286	0.34	0.645	0.83	0.877	0.96
2	250	20	$\{10, 40, \ldots, 250\}$	0.326	0.45	0.681	0.87	0.885	0.98
3	250	10	$\{10, 30, \ldots, 250\}$	0.232	0.26	0.487	0.54	0.683	0.90
4	250	20	$\{10, 30, \ldots, 250\}$	0.245	0.24	0.520	0.68	0.722	0.88
5	500	10	$\{10, 50, \ldots, 330\}$	0.264	0.48	0.621	0.84	0.765	1.0
6	500	20	$\{10, 50, \ldots, 330\}$	0.287	0.54	0.616	0.89	0.864	1.0
7	500	10	$\{10, 40, \ldots, 310\}$	0.232	0.29	0.516	0.75	0.710	0.95
8	500	20	$\{10, 40, \ldots, 310\}$	0.246	0.31	0.540	0.76	0.742	0.91

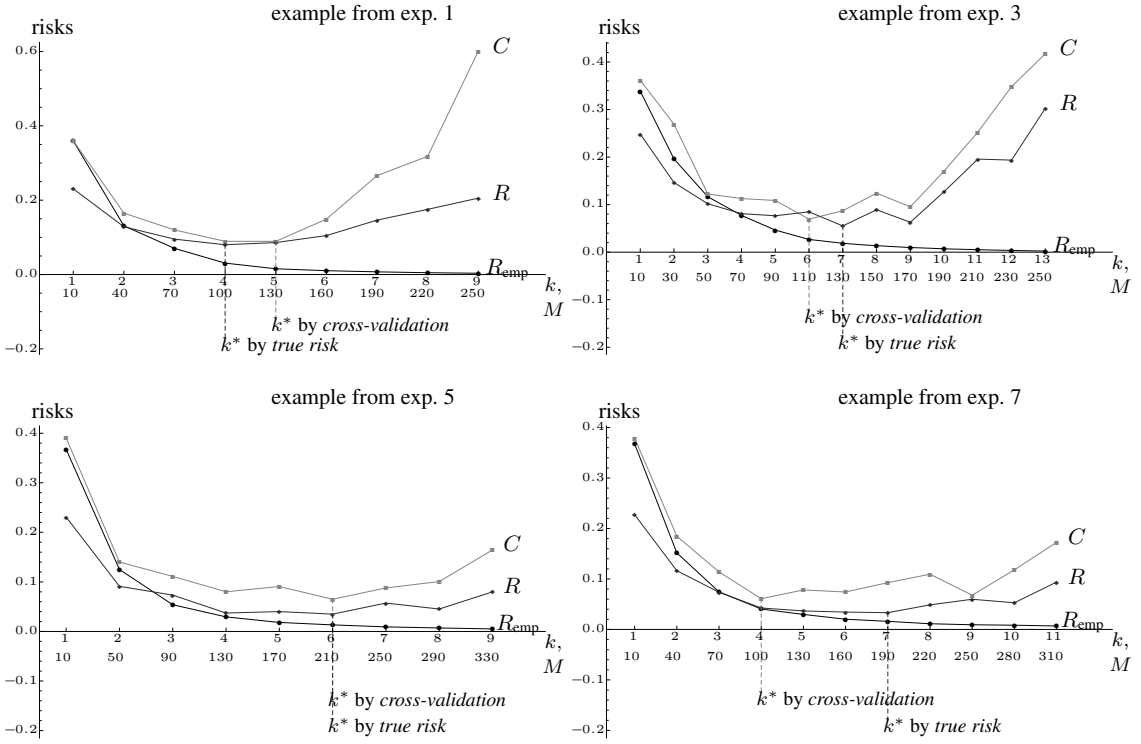

Fig. 5. Single examples of SRM procedures obtained in experiments for Scenario I. For each point k of the structure we present empirical risk $R_{\text{emp}}(\omega_{k,I})$, cross-validation result C_k and *true risk* $R(\omega_{k,I})$ calculated by an appropriate integral.

1. *What is the probability that the minimum point k^*, indicated via SRM, would agree with the minimum point indicated via cross-validation, if such a cross-validation was performed?*

2. *With what probability would the minimum point of unknown cross-validation results C_k fall into the neighbourhood of the minimum point k^* with a side Δ, indicated via SRM, if such cross-validation was performed?*

In other words, we save time by not performing cross-validation, i.e., we get $V_1, V_2, \ldots, V_K$ results $O(n)$ times faster, but we want to know to what extent our results are compliant with unknown results of cross-validation, if it was performed.

6.1. Distribution of cross-validation results. For the purpose of this scenario it would be useful to derive a distribution of type $C_k \sim N(V_k, \sigma)$ with a sufficiently tight σ. We state here the following theorem without a proof[15].

[15]In parallel with this publication a paper by Klęsk, in which this theorem is proved, undergoes a reviewing process in another journal.

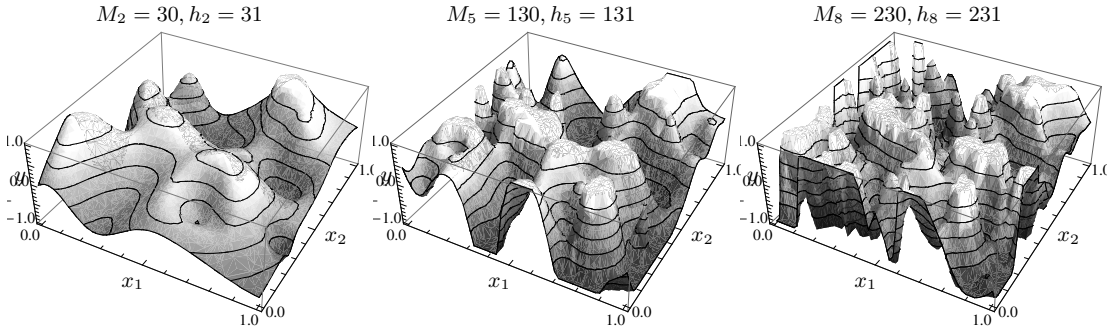

Fig. 6. Example models for regression estimation using a data set of size $I = 250$: $M_2 = 30 \rightarrow R(\omega_{2,I}) = 0.144$ (too simplistic), $M_5 = 130 \rightarrow R(\omega_{5,I}) = 0.058$ (duly complex—the best generalization), $M_8 = 230 \rightarrow R(\omega_{8,I}) = 0.257$ (too complex). Examples taken from the experiment no. 3.

Theorem 3. *For any $\eta > 0$, arbitrarily small, there is a small number*

$$\alpha(\eta, n) = \eta - \sum_{j=1}^{n} \binom{n}{j} (-1)^j (2\eta)^j \tag{65}$$

and the number

$$\begin{aligned} &\epsilon(\eta, I, N_k, n) \\ &\quad = B_k \left(2\sqrt{\frac{n}{n-1}} + 1 \right) \sqrt{\frac{\ln N_k - \ln \eta}{2I}} \\ &\qquad + B_k \left(\sqrt{n} + \sqrt{\frac{n}{n-1}} \right) \sqrt{\frac{-\ln \eta}{2I}}, \end{aligned} \tag{66}$$

such that

$$P\left(|V_k - C_k| \leq \epsilon(\eta, I, N_k, n) \right) \geq 1 - \alpha(\eta, n). \tag{67}$$

This theorem is useful, since we can compare its thesis with a suitable equality for normal distribution imposed on C_k (by means of the CLT):

$$\begin{aligned} P\left(|V_k - C_k| \leq \epsilon(\eta, I, N_k, n) \right) &\geq 1 - \alpha(\eta, n), \\ P\left(|V_k - C_k| \leq a_{1 - \frac{\alpha(\eta,n)}{2}} \sigma \right) &= 1 - \alpha(\eta, n). \end{aligned}$$

We see that we can pessimistically approximate each C_k as follows:

$$C_k \sim N\left(V_k, \frac{\epsilon(\eta, I, N_k, n)}{a_{1 - \frac{\alpha(\eta,n)}{2}}} \right). \tag{68}$$

6.2. Results of experiments. We impose analogical experimental conditions as in experiments for Scenario I: data sets, the set of approximating functions, the algorithm of the learning machine, etc. Results are shown in Table 2 and example illustrations are shown in Fig. 7.

Here are some comments on experimental results for this scenario:

1. The observed frequencies of the "agreement" between the minima of Vapnik bounds and cross-validation results are fairly high but generally smaller than for Scenario I, in which the agreement of cross-validation and true risks minima was compared.
2. If a practitioner were satisfied with the observed agreement frequency of about $75\% \div 90\%$ for a maximal discrepancy of $\Delta = 2$, he/she could save time by not performing n-fold cross-validation, just SRM, and obtain the model $O(n)$ times faster. Obviously, this comment is imprecise and valid only for experiments with similar conditions: the density of the structure, the proportion of the data set size to the capacity of the set of approximating functions.
3. Minimal probabilities estimated on the basis of the distribution (68) are much smaller than frequencies, which means that the distribution (68) is not sufficiently tight and, unfortunately, of little practical usefulness for accurate calculations.
4. Contrary to Scenario I, the 20-fold cross-validation made the calculated probabilities smaller than 10-fold cross-validation, which agrees with standard deviations in (68) and the influence of n therein.

7. Scenario III: Vapnik bounds and true risks

In this section we consider the SRM scenario again but we do not compare the obtained minimum to unknown

Table 2. Results of experiments for Scenario II. In each row there are shown the data set size, the structure, the calculated minimal probabilities and observed frequencies of the events considered. For all results $\eta = 0.05$, $\eta_B = 0.25$.

no.	I	n	number of terms M (successive points in the structure)	*Is unknown cross-validation minimum at point k^*?* min. P	*Is unknown cross-validation minimum at point k^*?* observed ν	*Is unknown cross-validation minimum in $U(k^*,1)$?* min. P	*Is unknown cross-validation minimum in $U(k^*,1)$?* observed ν	*Is unknown cross-validation minimum in $U(k^*,2)$?* min. P	*Is unknown cross-validation minimum in $U(k^*,2)$?* observed ν
1	2	3	4	5	6	7	8	9	10
1	250	10	$\{10, 40, \ldots, 250\}$	0.102	0.26	0.313	0.77	0.497	0.90
2	250	20	$\{10, 40, \ldots, 250\}$	0.110	0.24	0.272	0.61	0.448	0.82
3	250	10	$\{10, 30, \ldots, 250\}$	0.078	0.22	0.218	0.74	0.329	0.89
4	250	20	$\{10, 30, \ldots, 250\}$	0.060	0.18	0.190	0.59	0.301	0.79
5	500	10	$\{10, 50, \ldots, 330\}$	0.101	0.25	0.307	0.71	0.430	0.84
6	500	20	$\{10, 50, \ldots, 330\}$	0.105	0.22	0.268	0.55	0.468	0.75
7	500	10	$\{10, 40, \ldots, 310\}$	0.092	0.23	0.261	0.62	0.422	0.79
8	500	20	$\{10, 40, \ldots, 310\}$	0.089	0.19	0.247	0.45	0.426	0.75

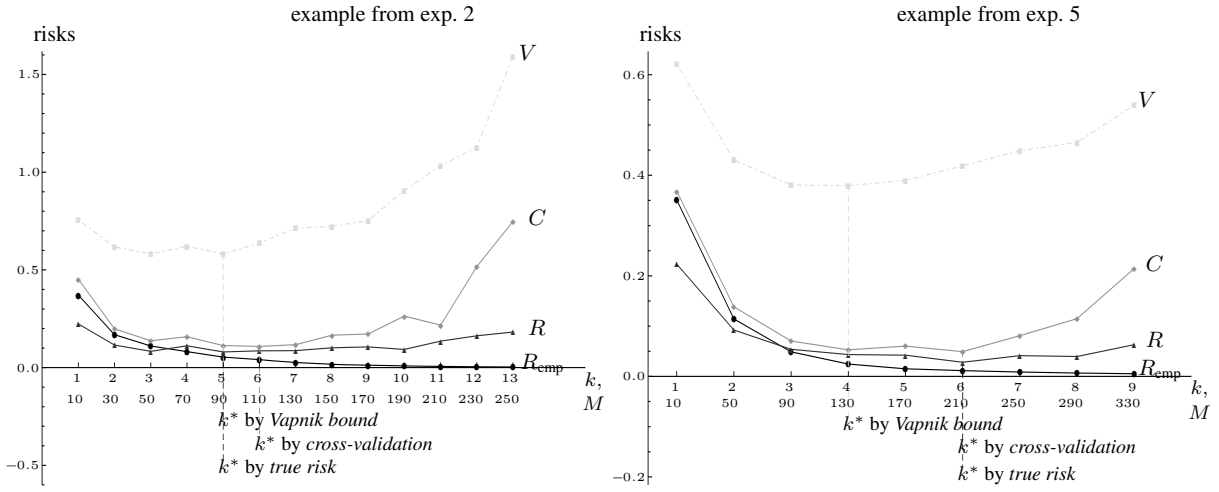

Fig. 7. Single examples of SRM procedures obtained in experiments for Scenario II. For each point k of the structure we give: empirical risk $R_{\text{emp}}(\omega_{k,I})$, cross-validation result C_k, Vapnik bound V_k and *true risk* $R(\omega_{k,I})$ calculated by an appropriate integral.

cross-validation results. We would like to compare it to the unknown minimum of true risks.

We iterate over the structure $S_1 \subset S_2 \subset \cdots \subset S_K$ and for each its subset S_k we do not perform cross-validation. We calculate only Vapnik's bound to true risk:

$$\overbrace{\int_{\mathbf{Z}} Q(\mathbf{z}, \omega_{k,I}) dF(\mathbf{z})}^{R(\omega_{k,I})} \leq \overbrace{\frac{1}{I}\sum_{i=1}^{I} Q(\mathbf{z}_i, \omega_{k,I})}^{R_{\text{emp}}(\omega_{k,I})} + B_k \sqrt{\frac{\ln N_k - \ln \eta}{2I}}, \qquad (69)$$

which holds true with probability at least $1 - \eta$. We shall denote by V_k the right-hand side of the bound.

When the procedure is finished for the whole structure, we have a sequence of results

$$V_1, V_2, \ldots, V_K,$$

and an *indication* that the optimal complexity is at a point k^*, such that

$$V_{k^*} = \min_{k \in \{1,\ldots,K\}} V_k. \qquad (70)$$

Therefore, as our final model we choose the function $f(\mathbf{x}, \omega_{k^*,I})$ for which the guaranteed true risk according to (69) is minimal.

We pose the following two questions, similarly as we did for the scenario with cross-validation:

1. *What is the probability that the point k^*, indicated via SRM, is truly the minimum point of all unknown true risks $R(\omega_{k,I})$?*

2. *With what probability does the true minimum of all unknown true risks $R(\omega_{k,I})$ fall into the neighbourhood of point k^*, indicated via SRM, with a side Δ?*

These questions are important, since answering them would tell us how reliable the result of complexity selection

indicated via SRM is for given conditions of the experiment.

In Scenario I, for a fixed position in the structure, we managed to derive the relationship like

$$R(\omega_I) \sim N\left(C, \frac{1}{\sqrt{n}}\sqrt{{\sigma_1}^2 + {\sigma_2}^2}\right),$$

to a large extent due to having at disposal *empirical testing risks* calculated as *means* in cross-validation folds, which in consequence led to normality by means of the CLT. In the current scenario the situation is more complicated. We would also like to find a relationship of type $R(\omega_I) \sim N(a, b)$, or actually with any distribution not necessarily normal, where a and b would be expressed in terms of relevant constants. But this time the only quantity we can rest on is $R_{\text{emp}}(\omega_I)$—the *empirical risk* calculated for the whole data set. We know that this quantity is in a sense *biased* and it does not tell us anything about the generalization error because the function $f(\omega_I)$ is chosen to minimize the error for the training set. In other words, we know that for a fixed data set of size I the value of $R_{\text{emp}}(\omega_I)$ gets smaller and smaller converging to zero with an increasing capacity of the set of functions.

Therefore, to derive *some* distribution density function, one could do the following:

- explicitly assume normality but with an increasing standard deviation along k,
- apply the generalization bound (69),
- put quantiles in agreement.

The density obtained by such an approach does not seem, however, credible nor useful to the author, mainly due to the biased $R_{\text{emp}}(\omega_I)$ and unjustified normality assumption. Therefore, we do not present any results for this scenario. We have to content ourselves with the generalization bound (63), which does not give a density, just a pessimistic quantile.

8. Summary

The complexity selection for the regression estimation learning task was taken under consideration, with the application of two known procedures: cross-validation and SRM. In both cases one encounters the problem of possible discrepancy between the minimum indicated empirically via a given procedure and the genuine minimum of true risks, which in practice is unknown. In the paper we pose this problem in a novel quantitative form, namely, as the problem of calculating *probabilities* of the occurrence of the discrepancy for given conditions of experiment. The solution to this problem can provide additional knowledge about an experiment and its uncertainty.

Three scenarios were discussed: (I) a comparison of the minimum indicated via cross-validation with the genuine minimum of true risks, (II) a comparison of the minimum indicated via Vapnik's bounds with an unknown minimum of cross-validation, (III) a comparison of the minimum indicated via Vapnik's bounds with the genuine minimum of true risks.

A greater focus was put on Scenario I, for which we stated and proved theorems enabling to calculate pessimistic probabilities of events:

- *where no discrepancy between minima occurs,*
- *where a discrepancy with a certain deviation* Δ *does occur.*

Main theorems (1) and (2) were stated in terms of all relevant constants: the sample size, the number of cross-validation folds, the capacity of the set of approximating functions, bounds on this set.

For given conditions of the structure, the most influential elements on the obtained output probabilities (24), (27), (59) are the capacity of the set of approximating functions, the sample size and the number of cross-validation folds. Looking at the densities (25) we derived, it can be seen that with respect to the sample size I the variance is scaled by factor $1/\sqrt{I}$, which is an intuitive probabilistic result, whereas with respect to the number of folds n the variance is scaled by an effective factor $1 + 1/\sqrt{n-1} + \sqrt{n}$. In particular, leave-one-out cross-validation strongly increases the variance. Clearly, to obtain high output probabilities (confirming the indicated point of the structure as the true minimum) one needs to provide a suitable combination of I and n which sufficiently tighten the variance. Apart from affecting output probabilities, these two factors also similarly affect the accuracy of normal approximations (CLT), which we show in Appendix B.

Empirical tests carried out for this scenario (Section 5) confirmed the results.

For Scenario II we stated a theorem (without proof) dedicated to the same purposes. However, we did not obtain satisfactory results in experiments (Section 6.2) due to an insufficiently tight standard deviation in the distribution $C_k \sim (V_k, \sigma)$, and pessimistic probabilities were not high enough. Still, we could draw some conclusions from the frequencies observed in experiments.

Looking generally at main theorems, two separate parts can be seen:

1. the approximation of probability densities which describe distributions of unknown true risks or cross-validation results (depending on the scenario),
2. the technique to calculate wanted probabilities by suitable integrals, using densities derived in the first part.

The second part alone—the calculation of probabilities—can be regarded as a solution to a more elementary problem not connected to machine learning itself, namely, the problem of finding the optimum of a *probabilistic function*[16] defined on a finite discrete set.

Acknowledgment

This work has been financed by the Polish Ministry of Science and Higher Education from the sources for science within the years 2010–2012 under Grant No. N N516 424938.

References

Anthony, M. and Shawe-Taylor, J. (1993). A result of Vapnik with applications, *Discrete Applied Mathematics* **48**(3): 207–217.

Bartlett, P. (1998). The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network, *IEEE Transactions on Information Theory* **44**(2): 525–536.

Bartlett, P., Kulkarni, S. and Posner, S. (1997). Covering numbers for real-valued function classes, *IEEE Transactions on Information Theory* **43**(5): 1721–1724.

Bartlett, P. and Tewari, A. (2007). Sample complexity of policy search with known dynamics, *Advances in Neural Information Processing Systems* **19**: 97–104.

Berry, A. (1941). The accuracy of the Gaussian approximation to the sum of independent variates, *Transactions of the American Mathematical Society* **49**(1): 122–136.

Bousquet, L., Boucheron S. and Lugosi G. (2004). *Introduction to Statistical Learning Theory*, Advanced Lectures in Machine Learning, Springer, Heidelberg, pp. 169–207.

Cherkassky, V. and Mulier, F. (1998). *Learning from Data*, John Wiley & Sons, Hoboken, NJ.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer, New York, NY.

Devroye, L., Gyorfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, NY.

Efron, B. and Tibshirani, R. (1993). *An Introduction to Bootstrap*, Chapman & Hall, London.

Esséen, C. (1942). On the Liapounoff limit of error in the theory of probability, *Arkiv fdr Matematik, Astronomi och Fysik* **28A**(9): 1–19.

Esséen, C. (1956). A moment inequality with an application to the central limit theorem, *Skand. Aktuarietidskr.* **39**: 160–170.

Fu, W., Caroll, R. and Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation, *Bioinformatics* **21**(9): 1979–1986.

[16]That is, values of the function are not deterministic but described by a certain probability density.

Graham, R., Knuth, D. and Patashnik, O. (2002). *Matematyka konkretna (Concrete Mathematics. A Foundation for Computer Science)*, PWN, Warsaw.

Hasterberg, T., Choi, N. H., Meier, L. and Fraley C. (2008). Least angle and l1 penalized regression: A review, *Statistics Surveys* **2**: 61–93.

Hellman, M. and Raviv, J. (1970). Probability of error, equivocation and the Chernoff bound, *IEEE Transactions on Information Theory* **16**(4): 368–372.

Hjorth, J. (1994). *Computer Intensive Statistical Methods Validation, Model Selection, and Bootstrap*, Chapman & Hall, London.

Knuth, D. (1997). *The Art of Computer Programming*, Addison-Wesley, Reading, MA.

Kohavi, R. (1995). A study of cross-validation and boostrap for accuracy estimation and model selection, *International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Quebec, Canada*, pp. 1137–1143.

Korzeń, M. and Klęsk, P. (2008). Maximal margin estimation with perceptron-like algorithm, *in* L. Rutkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh and J. Żurada (Eds.), *Artificial Intelligence and Soft Computing—ICAISC 2008*, Lecture Notes in Artificial Intelligence, Vol. 5097, Springer, Berlin, Heidelberg, pp. 597–608.

Krzyżak, A., Kohler M., and Schäfer D. (2000). Application of structural risk minimization to multivariate smoothing spline regression estimates, *Bernoulli* **8**(4): 475–489.

Ng, A. (2004). Feature selection, l_1 vs. l_2 regularization, and rotational invariance, *ACM International Conference on Machine Learning, Banff, Alberta, Canada*, Vol. 69, pp. 78–85.

Schmidt, J., Siegel, A. and Srinivasan, A. (1995). Chernoff-Hoeffding bounds for applications with limited independence, *SIAM Journal on Discrete Mathematics* **8**(2): 223–250.

Shawe-Taylor, J., Bartlett, P., Williamson, R. and Anthony, M. (1996). A framework for structural risk minimization, *COLT*, ACM Press, New York, NY, pp. 68–76.

Shevtsova, I. (2007). Sharpening of the upper bound of the absolute constant in the Berry–Esséen inequality, *Theory of Probability and its Applications* **51**(3): 549–553.

van Beek, P. (1972). An application of Fourier methods to the problem of sharpening the Berry–Esséen inequality, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **23**: 187–196.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer, New York, NY.

Vapnik, V. (1998). *Statistical Learning Theory: Inference from Small Samples*, Wiley, New York, NY.

Vapnik, V. (2006). *Estimation of Dependencies Based on Empirical Data*, Information Science & Statistics, Springer, New York, NY.

Vapnik, V. and Chervonenkis, A. (1968). On the uniform convergence of relative frequencies of events to their probabilities, *Doklady Akademii Nauk* **9**(4): 915–918.

Vapnik, V. and Chervonenkis, A. (1989). The necessary and sufficient conditions for the consistency of the method of empirical risk minimization, *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification and Forecasting*, Vol. 2, pp. 217–249.

Weiss, S. and Kulikowski, C. (1991). *Computer Systems That Learn*, Morgan Kauffman Publishers, San Francisco, CA.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes, *Journal of Machine Learning Research* **2**: 527–550.

Przemysław Klęsk received his M.Sc. and Ph.D. degrees at the Faculty of Computer Science, Technical University of Szczecin, Poland. At present, he works at the same University (currently named the Westpomeranian University of Technology) at the Department of Methods of Artificial Intelligence and Applied Mathematics as an assistant professor. His fields of interest are mathematics, data analysis and modeling, Vapnik's statistical learning theory, neural networks, genetic algorithms. He also has a relevant background in object-oriented programming, specifically in Java and J2EE technologies.

Appendix A

Justification of *minimal* probabilities in main theorems

The following theorem has significant meaning for the main Theorems (1 and 2) presented in the paper. It justifies the notion 'minimal probability' we used in those theorems—demonstrates that when improving (tightening) variances for any position in the structure the probability of interest also improves, which might not be intuitively obvious.

Theorem 4. *Let X_1 and X_2 be two independent normally distributed random variables with means $\mu_1 < \mu_2$ and variances σ_1 and σ_2, respectively. The probability that $X_1 < X_2$ can be calculated as*

$$\begin{aligned} &P(X_1 < X_2) \\ &\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [x_1 < x_2] p_1(x_1) p_2(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &\quad = \int_{-\infty}^{\infty}\left(\int_{x_1}^{\infty} p_2(x_2)\,\mathrm{d}x_2\right) p_1(x_1)\,\mathrm{d}x_1 \end{aligned} \tag{71}$$

or

$$P(X_1 < X_2) = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{x_2} p_1(x_1)\,\mathrm{d}x_1\right) p_2(x_2)\,\mathrm{d}x_2, \tag{72}$$

where p_1, p_2 denote normal density functions for X_1 and X_2, respectively. Suppose now that we are able to tighten either variance, i.e., to introduce $0 < \sigma_1' \leq \sigma_1$ and $0 < \sigma_2' \leq \sigma_2$, and thus create new random variables: $X_1' \sim N(\mu_1, \sigma_1')$, $X_2' \sim N(\mu_2, \sigma_2')$ but with original means.

Then, for any choice of σ_1', σ_2', the probability $P(X_1' < X_2')$ is not worse than $P(X_1 < X_2)$:

$$\forall_{\substack{0<\sigma_1'\leq\sigma_1 \\ 0<\sigma_2'\leq\sigma_2}} \quad P(X_1' < X_2') \geq P(X_1 < X_2). \tag{73}$$

Proof. First, we analyze the case when $\sigma_1' = \sigma_1$ is fixed whereas σ_2' is relaxed and can be freely tightened: $0 < \sigma_2' \leq \sigma_2$. Consider the function

$$\delta(x_1) = \int_{x_1}^{\infty} p_2'(x_2)\,\mathrm{d}x_2 - \int_{x_1}^{\infty} p_2(x_2)\,\mathrm{d}x_2. \tag{74}$$

We shall call it the *transfer* function, since for given x_1 it represents the amount of the probability measure *transfered* from the interval $(-\infty, x_1)$ into the interval $[x_1, \infty)$ in the inner integral of (71) when the switch from p_2 to p_2' is made. One can observe that $\delta(x_1) > 0$ for $-\infty < x_1 < \mu_2$ (since the left tail excluded from the integration is heavier for p_2 than for p_2'), in the middle $\delta(\mu_2) = \frac{1}{2} - \frac{1}{2} = 0$, and finally $\delta(x_1) < 0$ for $\mu_2 < x_1 < \infty$. Moreover, $\delta(x_1)$ first increases until it reaches a point corresponding to the first intersection of p_2 and p_2', next it decreases until it reaches the second intersection point, and then it increases again asymptotically to 0, see Fig. 8.

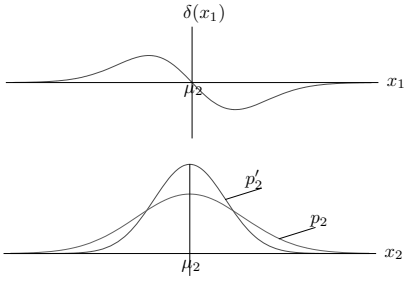

Fig. 8. Illustration of the *transfer* function δ.

We remark that when $\delta > 0$, it potentially favors the *increase* of the probability $P(X_1' < X_2')$ with respect to $P(X_1 < X_2)$, since realizations of $X_2' > x_1$ are then more likely than originally due to the transfer of the probability measure. On the other hand, when $\delta < 0$, it potentially favors a *decrease* in the probability $P(X_1' < X_2')$ with respect to $P(X_1 < X_2)$, since realizations of $X_2' > x_1$ are then less probable than originally due to transfer of the probability measure. We need to check which situation of

the two is more frequent.

$$\begin{aligned}
&P(X_1' < X_2') - P(X_1 < X_2) \\
&\quad = \int_{-\infty}^{\infty} \left(\int_{x_1}^{\infty} p_2'(x_2)\,\mathrm{d}x_2 \right) p_1(x_1)\,\mathrm{d}x_1 \\
&\qquad - \int_{-\infty}^{\infty} \left(\int_{x_1}^{\infty} p_2(x_2)\,\mathrm{d}x_2 \right) p_1(x_1)\,\mathrm{d}x_1 \\
&\quad = \int_{-\infty}^{\infty} \Bigg(\underbrace{\int_{x_1}^{\infty} \Big(p_2'(x_2) - p_2(x_2) \Big)\,\mathrm{d}x_2}_{\delta(x_1)} \Bigg) p_1(x_1)\,\mathrm{d}x_1 \\
&\quad = \int_{-\infty}^{\mu_2} \underbrace{\delta(x_1)}_{\geq 0} p_1(x_1)\,\mathrm{d}x_1 + \int_{\mu_2}^{\infty} \underbrace{\delta(x_1)}_{\leq 0} p_1(x_1)\,\mathrm{d}x_1 > 0.
\end{aligned} \tag{75}$$

The last inequality is true since $\mu_2 > \mu_1$ and therefore the first integral is heavier (i.e., it contains more probability measure of X_1) than the second:

$$\int_{-\infty}^{\mu_2} |\delta(x_1)| p(x_1)\,\mathrm{d}x_1 > \int_{\mu_2}^{\infty} |\delta(x_1)| p(x_1)\,\mathrm{d}x_1$$

while we keep in mind that δ is symmetric with respect to the point μ_2.

The analysis for the second case when $\sigma_2' = \sigma_2$ is fixed whereas σ_1' can be freely tightened is analogical. It is sufficient to change the order of integration, i.e., to use the formula (72) instead of (71) and to redefine δ as

$$\delta(x_2) = \int_{-\infty}^{x_2} p_1'(x_1)\,\mathrm{d}x_1 - \int_{-\infty}^{x_2} p_1(x_1)\,\mathrm{d}x_1.$$

Having proved both cases: (a) tightening σ_1 (with σ_2 fixed) does not worsen the target probability, (b) tightening σ_2 (with σ_1 fixed) does not worsen the target probability, implies that for any sequence of actions which tighten first σ_1 then σ_2 or vice-versa the target probability does not worsen. Hence, for *any* σ_1', σ_2', $P(X_1' < X_2') \geq P(X_1 < X_2)$ holds true. ■

In Theorems 1 and 2 we approximated cross-validation results C_k along the structure by normal distributions (CLT) and assumed the pessimistic variances for them. This fact is expressed in (33), (48) and (51). Now, owing to Theorem 4, it is *guaranteed* that by improving (tightening) variances for *any* position in the structure we also improve the minimal probabilities (24), (27) asserted in the main theorems, we certainly do not worsen them.

Appendix B

Accuracy of normal approximations

For informative purposes we give several remarks on the accuracy of our normal approximations on the basis of the Berry–Esséen theorem (Berry, 1941; Essèen, 1942;1956)

Suppose that a sequence of cumulative distribution functions F_i converges, as $i \to \infty$, to some target F. This is a weak convergence statement and it says nothing about the accuracy of the approximation for a particular finite value of i. We need to have an idea of the committed error, i.e., $|F_i(x) - F(x)|$ (DasGupta, 2008). Owing to CLT, for a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_I$, it is known that

$$\frac{\overline{X_I} - E(\overline{X_I})}{\sqrt{\mathrm{Var}(\overline{X_I})}} \underset{I\to\infty}{\longrightarrow} Z \sim N(0,1). \tag{76}$$

Different results giving a bound on the approximation error for finite I typically make assumptions about moments of X_i (DasGupta, 2008). A classical result of this kind is the following theorem.

Theorem 5. (Berry–Esséen) *Let $X_1, \ldots, X_I$ be i.i.d. with $E(X_i) = \mu$, Var$(X_i) = \sigma^2$, and $\beta_3 = E(|X_i - \mu|^3) < \infty$. Then there exists a universal constant C, not depending on I or the distribution of the X_i, such that*

$$\sup_x \left| P\left(\frac{\sqrt{I}(\overline{X} - \mu)}{\sigma} \leq x \right) - \Phi(x) \right| \leq \frac{C\beta_3}{\sigma^3 \sqrt{I}}, \tag{77}$$

where $\Phi(x)$ is the cumulative distribution function for $N(0,1)$.

The universal constant C has been refined historically from $C = 0.7975$ by van Beek (1972) to $C = 0.7056$ by Shevtsova (2007).

As one can see in the theorem, the uniform error decreases with $1/\sqrt{I}$ asymptotically to 0. Looking back closely at Theorems 1 and 2 and the formulas (25), (26) presented in this paper, we see that the actual number of random variables we sum up is implied both by the sample size I and the number of cross-validation folds n. It can be noted that the effective number of summands (repetitions) is

$$\frac{I}{\left(1 + \frac{1}{\sqrt{n-1}} + \sqrt{n}\right)^2},$$

and this is the number that should be inserted into the estimate from the Berry–Esséen theorem rather than $1/\sqrt{I}$ alone. Therefore, the uniform error between cumulative distribution functions for unknown true risks and their normal approximations that we propose is at most

$$\frac{C\beta_3}{\sigma^3} \left(1 + \frac{1}{\sqrt{n-1}} + \sqrt{n}\right) \frac{1}{\sqrt{I}}, \tag{78}$$

where σ and β_3 denote respectively the standard deviation and the third moment of the unknown distribution we approximate.

It can be checked that the influence of the sample size I on the error is naturally favorable, i.e., the error

decreases with the square root of the sample size, whereas the influence of the number of cross-validation folds n is unfavorable. One can realize it by looking at the most extreme case of leave-one-out cross-validation which corresponds to $n = I$. Then, the variance of testing errors per fold, which depends on $1/\sqrt{I/n}$, is the largest. We illustrate the influences of I and n on the error in Fig. 9.

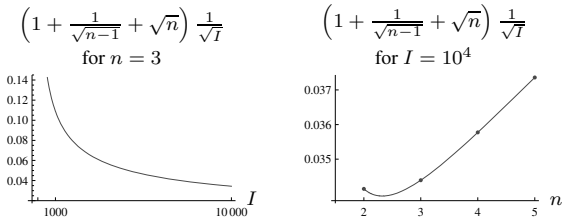

Fig. 9. Illustration of the influence of the sample size I and the number of cross-validation folds on the error of normal approximation.

Interestingly, it can be checked that there is a minimum point of error with respect to the number of cross-validation folds at

$$n^* = 1 + \frac{1}{3}\left(\frac{27}{2} - \frac{3\sqrt{69}}{2}\right)^{1/3} + \frac{\left(\frac{1}{2}(9+\sqrt{69})\right)^{1/3}}{3^{2/3}} \approx 2.33$$

if n is treated as a continuous variable.

In general, unfortunately nothing reasonable can be said about σ, β_3 (and their ratio) for unknown distributions we approximate. Their values depend in a peculiar way on two factors: (1) the joint probability density $p(\mathbf{z})$ representing the specific learning problem and (2) properties of the learning machine (how it selects the output function and what the properties of that function like smoothness, boundedness, etc. are). In a very particular case, if approximated unknown distributions were in fact normal distributions, then it could be checked that $\lim_{\sigma\to\infty} \beta_3/\sigma^3 = 2\sqrt{2/\pi} \approx 1.596$. It would give us an approximate idea about the value of ratio $C\beta_3/\sigma \approx 1.126$, using the value for $C = 0.7056$ by Shevtsova (2007). But this assumption clearly does not have to be true in the general case.

Skipping unknown constants, we state in the theorems the *order* of the uniform error, with respect to I and n alone, to be

$$O\left(\left(1+\frac{1}{\sqrt{n-1}}+\sqrt{n}\right)\frac{1}{\sqrt{I}}\right), \tag{79}$$

If we want to make this order less than a certain small ϵ, then we need to provide for the experiment a sample of size

$$I > \frac{1}{\epsilon^2}\left(1+\frac{1}{\sqrt{n-1}}+\sqrt{n}\right)^2,$$

given a fixed n. For example, for $\epsilon = 0.01$ and $n = 3$, a sample of size $I \approx 1.2 \cdot 10^5$ is required, for $\epsilon = 0.05$, $n = 3$, it is sufficient to have $I \approx 4.8 \cdot 10^3$. In the experiments shown in Section 5 we set up even more relaxed conditions due to computational costs[17]: $I = 500$, $n = 10$, and, as Table 1 shows, the results are still satisfactory.

Appendix C

Expected square deviation for a nested random variable

Lemma 3. *Let X_1 be a random variable with the expected value EX_1 and the variance ${\sigma_{X_1}}^2$. Assume that for each realization x_1 of X_1 there is another (nested) random variable $X_2|x_1$ with the following expectation and variance:*

$$\forall x_1 \quad E(X_2|x_1) = x_1, \qquad D({X_2}^2|x_1) = {\sigma_{X_2}}^2 = const..$$

Then the expected square deviation of X_2 from EX_1 is ${\sigma_{X_1}}^2 + {\sigma_{X_2}}^2$.

Proof.

$$\begin{aligned}
&\int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty}(x_2-EX_1)^2 p_{X_2|x_1}(x_2)\,\mathrm{d}x_2\Big)p_{X_1}(x_1)\,\mathrm{d}x_1\\
&= \int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty}(x_2-x_1+x_1-EX_1)^2\\
&\qquad \times p_{X_2|x_1}(x_2)\,\mathrm{d}x_2\Big)p_{X_1}(x_1)\,\mathrm{d}x_1\\
&= \int_{-\infty}^{\infty}\Bigg(\int_{-\infty}^{\infty}(x_2-x_1)^2 p_{X_2|x_1}(x_2)\,\mathrm{d}x_2\\
&\qquad + 2(x_1-EX_1)\underbrace{\int_{-\infty}^{\infty}(x_2-x_1)p_{X_2|x_1}(x_2)\,\mathrm{d}x_2}_{0}\\
&\qquad + (x_1-EX_1)^2\int_{-\infty}^{\infty}p_{X_2|x_1}(x_2)\,\mathrm{d}x_2\Bigg)\\
&\qquad \times p_{X_1}(x_1)\,\mathrm{d}x_1\\
&= \int_{-\infty}^{\infty}\left({\sigma_{X_2}}^2 + (x_1-EX_1)^2\right)p_{X_1}(x_1)\,\mathrm{d}x_1\\
&= {\sigma_{X_2}}^2 + {\sigma_{X_1}}^2.
\end{aligned}$$

■

Received: 19 June 2009
Revised: 16 March 2010
Re-revised: 19 March 2010

[17] Each experiment was carried out with 100 repetitions.