A RELATIONSHIP BETWEEN CROSS-VALIDATION AND VAPNIK BOUNDS ON GENERALIZATION OF LEARNING MACHINES

Przemysław Klęsk
Department of Methods of Artificial Intelligence and Applied Mathematics, Westpomeranian University of Technology
ul. Żolnierska 49, Szczecin, Poland

Keywords: Statistical learning theory, Bounds on generalization, Cross-validation, Empirical risk minimization, Structural risk minimization, Vapnik–Chervonenkis dimension.

Abstract: Typically, the n-fold cross-validation is used both to: (1) estimate the generalization properties of a model of fixed complexity, (2) choose from a family of models of different complexities, the one with the best complexity, given a data set of certain size. Obviously, it is a time-consuming procedure. A different approach — the *Structural Risk Minimization* is based on generalization bounds of learning machines given by Vapnik (Vapnik, 1995a; Vapnik, 1995b). Roughly speaking, SRM is $O(n)$ times faster than n-fold cross-validation but less accurate.

We state and prove theorems, which show the probabilistic relationship between the two approaches. In particular, we show what ε-difference between the two, one may expect without actually performing the cross-validation. We conclude the paper with results of experiments confronting the probabilistic bounds we derived.

1 INTRODUCTION AND NOTATION

One part of the *Statistical Learning Theory* developed by Vapnik (Vapnik, 1995a; Vapnik, 1995b; Vapnik, 2006) is the *theory of bounds*. It provides probabilistic bounds on generalization of learning machines. The key mathematical tools applied to derive the bounds in their additive versions are Chernoff and Hoeffding inequalities[1] (Vapnik, 1995b; Cherkassky and Mulier, 1998; Hellman and Raviv, 1970; Schmidt et al., 1995).

We use this theory to show a probabilistic relationship between two approaches for complexity selection: *n-fold cross-validation* (popular among practioner modelers) and *Structural Risk Minimization* proposed by Vapnik (rarely met in practice) (Shawe-Taylor et al., 1996; Devroye et al., 1996; Anthony and Shawe-Taylor, 1993; Krzyżak et al., 2000). We remind that SRM is $O(n)$ times faster than n-fold cross-validation (since SRM does not perform any repetitions/folds per single fixed complexity, nor testing) but less accurate, since the selection of optimal complexity is based on the guaranteed generalization risk. The bound for the guaranteed risk is expressed in terms of *Vapnik-Chervonenkis dimension*, and is a pessimistic overestimation of the *growth function*, which in turn is overestimation of the unknown *Vapnik-Chervonenkis entropy*. We formally remind these notions later in the paper. All those overestimations contribute (unfortunately) to the fact that for a fixed sample size, SRM usually underestimates the optimal complexity and chooses too simple model.

Results presented in this paper may be regarded as conceptually akin to results by Holden (Holden, 1996a; Holden, 1996b), where error bounds on cross-validation and so-called *sanity-check* bounds are derived. The sanity-check bound is a proof, for large class of learning algorithms, that the error of the *leave-one-out* estimate is not much worse — $O(\sqrt{h/I})$ — than the worst-case behavior of the training error estimate, where h stands for Vapnik-Chervonenkis dimension of given set of functions and I stands for the sample size. The name sanity-check refers to the fact that although we believe that under many circumstances, the leave-one-out estimate will

[1] Chernoff inequality: $P\big(|\nu_I - p| \geq \varepsilon\big) \leq 2\exp(-2\varepsilon^2 I)$, Hoeffding inequality: $P\big(|X_I - EX| \geq \varepsilon\big) \leq 2\exp(-\frac{2\varepsilon^2 I}{B^2 - A^2})$. Meaning (respectively): observed frequencies on a sample of size I converge to the true probability as I grows large; analogically: means of a random variable (bounded by A and B) converge to the expected value. It is a *in-probability-convergence* and its rate is exponential.

Klęsk P..
A RELATIONSHIP BETWEEN CROSS-VALIDATION AND VAPNIK BOUNDS ON GENERALIZATION OF LEARNING MACHINES.
DOI: 10.5220/0003121000050017
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 5-17
ISBN: 978-989-8425-40-9

perform better than the training error (and thus justify its computational expense) the goal of the sanity-check bound is to simply prove that it is not much worse than the training error (Kearns and Ron, 1999).

These results were further generalized by Kearns (Kearns and Ron, 1999; Kearns, 1995a; Kearns, 1995b) using the notion of (β_1,β_2)-*error stability*[2] rather than (β_1,β_2)-*hypothesis stability*[3] imposed on the learning algorithm.

For the sake of comparison and to set up the perspective for further reading of this paper, we highlight some differences of meaning of our results and the results mentioned above:

- we do not focus on how well the *measured* cross-validation result estimates the generalization error or how far it is from the training error in the leave-one-out case — sanity-check bounds (Holden, 1996b; Kearns and Ron, 1999); instead, we want to make statements about the cross-validation result *without actually measuring it*, thus, remaining in the setting of the SRM framework.
- in particular we want to state probabilistically what ε-difference one can expect between the *known* Vapnik bound and the *unknown* cross-validation result for given conditions of the experiment,
- in the consequence, we want to be able to calculate the necessary size of the training sample, so that the ε is sufficiently small, and so that the optimal complexity indicated via SRM is acceptable in the sense that cross-validation, if performed, would probably indicate the same complexity; this statement may seem related to the notion of *sample complexity* considered e.g. by Bartlett (Bartlett et al., 1997; Bartlett, 1997) or Ng (Ng, 2004), but we do not find the sample size required for the algorithm to learn/generalize "well" but rather such a sample size so that complexity selection via SRM gives similar results to complexity selection via cross-validation,
- we do not explicitly introduce the notion of error stability for the learning algorithm, but this kind of stability is implicitly derived be means of Chernoff-Hoeffding-like inequalities we write.
- we do not focus on the leave-one-out cross-validation; we consider a more general *n-fold non-stratified cross-validation* (also: more convenient for our purposes); the leave-one-out case can be read out from our results as a special case.

[2]We say that a learning algorithm has a (β_1,β_2)-error stability, if generalization errors for two models provided by this algorithm using respectively a training sample of size I and a sample with size lowered to $I-1$ are β_1-close to each other with probability at least $1-\beta_2$. Obviously the smaller both β_1,β_2 are the more stable the algorithm.

[3]We say that a learning algorithm has a (β_1,β_2)-hypothesis stability, if the two models provided by this algorithm using respectively a training sample of size I and sample with size lowered to $I-1$ are β_1-close to each other with probability at least $1-\beta_2$, where closeness of models is measured by some functional metrics, e.g. L_1, L_2, etc.

1.1 Notation Related to Statistical Learning Theory

We keep the notation similar to Vapnik's (Vapnik, 1995b; Vapnik, 1995a).

- 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 shortly 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, y_i are output values corresponding to them, and I is the set size. y_i differ depending on the learning task: for *classification* (pattern-recognition) $y_i\in\{1,2,\ldots,K\}$ — finite discrete set, for *regression estimation* $y_i\in\mathbb{R}$.

- We denote the *set of approximating functions* (models) in the sense of both classification or regression estimation as:

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

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

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

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

where $p(\mathbf{x})$ is the distribution density of input points, $p(y|\mathbf{x})$ is the 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 distribution density for pairs $(\mathbf{x},y)$. In practice, $p(\mathbf{x},y)$ is unknown but *fixed*, and hence we assume the sample $\{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_I\}$ to be *i.i.d.*[4] L is the so called *loss function* which measures the discrepancy between the output y and the model f. For classification, L is an indicator function:

$$L\Big(f(\mathbf{x},\omega),y\Big)=\begin{cases}0, & \text{for } y=f(\mathbf{x},\omega);\\ 1, & \text{for } y\neq f(\mathbf{x},\omega),\end{cases} \tag{2}$$

[4]Independent, identically distributed.

and the risk functional becomes $R(\omega) = \int_{\mathbf{x}\in\mathbf{X}} \sum_{y\in Y} L\big(f(\mathbf{x},\omega),y\big)\, p(\mathbf{x},y)\mathbf{dx}$. For regression estimation, L is usually chosen as the distance in L_2 metric:

$$L\Big(f(\mathbf{x},\omega),y\Big) = \big(f(\mathbf{x},\omega)-y\big)^2, \tag{3}$$

and the risk functional becomes $R(\omega) = \int_{\mathbf{x}\in\mathbf{X}} \int_{y\in Y} \big(f(\mathbf{x},\omega)-y\big)^2 p(\mathbf{x},y)dy\mathbf{dx}$.

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

$$R(\omega_0) = \inf_{\omega\in\Omega} R(\omega). \tag{4}$$

- Since only a finite set of samples $\{\mathbf{z}_1,\ldots,\mathbf{z}_I\}$ is at 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*:

$$R_{\text{emp}}(\omega) = \frac{1}{I}\sum_{i=1}^{I} L(y_i, f(\mathbf{x}_i,\omega)), \tag{5}$$

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

$$R_{\text{emp}}(\omega_I) = \inf_{\omega\in\Omega} R_{\text{emp}}(\omega) \tag{6}$$

— *Empirical Risk Minimization* principle (Vapnik, 1995a; Vapnik, 1995b; Vapnik and Chervonenkis, 1989; Cherkassky and Mulier, 1998).

- For simplification of notation and further considerations, we introduce 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[5] $\{f(\mathbf{x},\omega)\}_{\omega\in\Omega}$, we equivalently consider the *set of error functions* $\{Q(\mathbf{z},\omega)\}_{\omega\in\Omega}$. It is a 1:1 correspondence [6]. Now, we write the true risk as:

$$R(\omega) = \int_{\mathbf{z}\in\mathbf{X}\times Y} Q(\mathbf{z},\omega)\underbrace{p(\mathbf{z})}_{p(\mathbf{x},y)}\mathbf{dz} = \int_{\mathbf{Z}} Q(\mathbf{z},\omega)dF(\mathbf{z}), \tag{7}$$

and the empirical risk as

$$R_{\text{emp}}(\omega) = \frac{1}{I}\sum_{i=1}^{I} Q(\mathbf{z}_i,\omega)), \tag{8}$$

[5] In the sense of all learning tasks.

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

1.2 Notation Related to Cross-validation

In the paper, we shall consider the *non-stratified* variant of the n-fold cross-validation procedure (Kohavi, 1995). In each single fold (iteration) we first permute the data set and then we split it at the same fixed point into two disjoint subsets — a training set and a testing set. Thus, we guarantee the randomness by permutation per each fold, and among folds we do not care to make training sets disjoint pairwise. Since permutations are independent, hence *folds are independent* as well.

Such an approach is somewhere in-between the classical n-fold cross-validation and the *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) and hence folds are dependent, whereas in the bootstrapping instead of repeatedly analyzing subsets of data set, one repeatedly analyzes the subsamples (with replacement) of the data. For more information see (Hjorth, 1994; Weiss and Kulikowski, 1991; Fu et al., 2005).

We introduce the following notation. I' and I'' stand for the size of training and testing sets respectively.

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

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

In a single fold, let

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

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

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

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

represent respectively the training error and the testing error, calculated for any function ω.

By $\omega_{I'}$ we define the function that minimizes the *empirical training risk*

$$R'_{\text{emp}}(\omega_{I'}) = \inf_{\omega\in\Omega} R'_{\text{emp}}(\omega) \tag{11}$$

when the context of discussion is constrained to single fold. When, we will need to broaden the context onto

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

$$C = \frac{1}{n}\sum_{k=1}^{n} R''_{\text{emp}}(\omega_{I',k}). \tag{12}$$

The *independence of folds* can be formally expressed in the following way. For any two indices of folds $k \neq l$ and for any numbers A, B:

$$P(R''_{\text{emp}}(\omega_{I',k}) = A, R''_{\text{emp}}(\omega_{I',l}) = B) = P(R''_{\text{emp}}(\omega_{I',k}) = A) \cdot P(R''_{\text{emp}}(\omega_{I',l}) = B).$$

We stress the independence once again, because later on we are going to sum up several independent probabilistic inequalities into one inequality, and we would like the result to be true with the effective probability being the product of component probabilities.

2 THE RELATIONSHIP FOR A FINITE SET OF APPROXIMATING FUNCTIONS

2.1 Classification Learning Task

Similarly to Vapnik, let us start with the classification learning task and the simplest case of a *finite* set of N indicator functions: $\{Q(\mathbf{z},\omega_j)\}_{\omega_j\in\Omega}$, $j = 1,2,\ldots,N$. Not to complicate things, we will keep on writing ω_I in the sense of the optimal function minimizing the empirical risk on our finite sample of size I, instead of writing more formally e.g. $\omega_{j(I)}$ [7].

Vapnik shows (Vapnik, 1995a; Vapnik, 1995b) that with probability at least $1-\eta$, the following bound on the true risk is satisfied:

$$\underbrace{\int_{\mathbf{Z}} Q(\mathbf{z},\omega_I)dF(\mathbf{z})}_{R(\omega_I)} \leq \underbrace{\frac{1}{I}\sum_{i=1}^{I} Q(\mathbf{z}_i,\omega_I)}_{R_{\text{emp}}(\omega_I)} + \sqrt{\frac{\ln N - \ln\eta}{2I}}. \tag{13}$$

The argument is the following:

$$P\Big(\sup_{1\leq j\leq N} R(\omega_j) - R_{\text{emp}}(\omega_j) \geq \varepsilon\Big) \leq \sum_{j=1}^{N} P\Big(R(\omega_j) - R_{\text{emp}}(\omega_j) \geq \varepsilon\Big) \leq N\cdot\exp(-2\varepsilon^2 I).$$

[7] In the sense that $j(I) \in \{1,\ldots,N\}$ returns the index of the minimizer given our data set of size I.

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

$$R(\omega_j) - R_{\text{emp}}(\omega_j) \leq \sqrt{\frac{\ln N - \ln\eta}{2I}},$$

which holds true with probability at least $1-\eta$ simultaneously *for all* functions in the set, since it holds for the worst. Hence, in particular it holds true for the function ω_I. And one gets the bound (13).

For the theorems to follow, we denote the right-hand-side in the Vapnik bound by $V = R_{\text{emp}}(\omega_I) + \sqrt{(\ln N - \ln\eta)/(2I)}$.

Theorem 1. *Let $\{Q(\mathbf{z},\omega_j)\}_{\omega_j\in\Omega}$, $j = 1,2,\ldots,N$, be a finite set of indicator functions (classification task) of size N. Then, for any $\eta > 0$, arbitrarily small, there is a small number*

$$\alpha(\eta,n) = \eta - \sum_{k=1}^{n}\binom{n}{k}(-1)^k(2\eta)^k, \tag{14}$$

and the number

$$\varepsilon(\eta,I,N,n) = \left(2\sqrt{\frac{n}{n-1}}+1\right)\sqrt{\frac{\ln N - \ln\eta}{2I}} + \left(\sqrt{n}+\sqrt{\frac{n}{n-1}}\right)\sqrt{\frac{-\ln\eta}{2I}}, \tag{15}$$

such that:

$$P\Big(|V - C| \leq \varepsilon(\eta,I,N,n)\Big) \geq 1 - \alpha(\eta,n). \tag{16}$$

Before we prove theorem 1, the following two remarks should be clear.

Remark 1. *The value of $\alpha(\eta,n)$ is monotonous with η. I.e. the smaller η we choose, the smaller $\alpha(\eta,n)$ becomes as well. Therefore the minimum probability measure $1-\alpha(\eta,n)$ is suitably large.*

$$\lim_{\eta\to 0^+}\left(\eta - \sum_{k=1}^{n}\binom{n}{k}(-1)^k(2\eta)^k\right) = \lim_{\eta\to 0^+}\left(\eta + 1 - \sum_{k=0}^{n}\binom{n}{k}(-1)^k(2\eta)^k\right) = \lim_{\eta\to 0^+}\left(\eta + 1 - \underbrace{(1-2\eta)^n}_{\to 1}\right) = 0.$$

Remark 2. *For the fixed values of η, N, n, the value of $\varepsilon(\eta,I,N,n)$ converges to zero as the sample size I grows large.*

This is an important remark, because it means that both the cross-validation result C and the Vapnik bound V *converge in probability*[8] to the same value[9] as the sample size grows large. Moreover, the rate of this convergence is exponential.

Proof of Remark 2. Since N is fixed, we note that for $\eta \to 0^+$

$$\sqrt{\frac{\ln N - \ln \eta}{2I}} \sim \sqrt{\frac{-\ln \eta}{2I}}.$$

Therefore, for fixed η, N, n there exists a constant, say D, such that

$$\varepsilon(\eta, I, N, n) = 2\left(\sqrt{\frac{n}{n-1}} + 1\right)\sqrt{\frac{\ln N - \ln \eta}{2I}} + \left(\sqrt{n} + \sqrt{\frac{n}{n-1}}\right)\sqrt{\frac{-\ln \eta}{2I}} \leq D\sqrt{\frac{-\ln \eta}{2I}}.$$

Solving the inequality for η we obtain $\eta \leq \exp(-2I\varepsilon^2/D^2)$. □

Having in mind the inequality (16), we now give two theorems in which the absolute value sign in $|V - C|$ is omitted. They can be viewed as the *upper* and the *lower* probabilistic bounds on C and they are derived as tighter bounds than (16). Proving these two theorems immediately implies proving the theorem 1.

Theorem 2. *With probability* $1 - \alpha(\eta, n)$ *or greater, the following inequality holds true:*

$$C - V \leq \left(\sqrt{\frac{n}{n-1}} - 1\right)\sqrt{\frac{\ln N - \ln \eta}{2I}} + \left(\sqrt{n} + \sqrt{\frac{n}{n-1}}\right)\sqrt{\frac{-\ln \eta}{2I}}. \quad (17)$$

Theorem 3. *With probability* $1 - \alpha(\eta, n)$ *or greater, the following inequality holds true:*

$$V - C \leq \left(2\sqrt{\frac{n}{n-1}} + 1\right)\sqrt{\frac{\ln N - \ln \eta}{2I}} + \sqrt{n}\sqrt{\frac{-\ln \eta}{2I}}. \quad (18)$$

The second result is more interesting, provided of course that the bound is positive for given constants η, I, N, n. Otherwise, we get zero or negative bound, which is trivial. The fig. 1 illustrates the sense of theorems 2 and 3.

[8] We say that $A(I)$ *converges in probability* to B, we write $A(I) \xrightarrow[I\to\infty]{P} B$, when for any numbers $\varepsilon > 0$, $\eta > 0$, there exists a treshold size of sample $I(\varepsilon, \eta)$, such that *for all* $I \geq I(\varepsilon, \eta)$: $P\big(|A(I) - B| > \varepsilon\big) \leq \eta$.

[9] C and V can be viewed as random variables, due to random realizations of data set $\{\mathbf{z}_1, \ldots, \mathbf{z}_I\}$ with joint density $p(\mathbf{z})$ (this affects C and V) and due to random realizations of subsets in cross-validation folds (this affects C). When the data set $\{\mathbf{z}_1, \ldots, \mathbf{z}_I\}$ is fixed, V is fixed too.

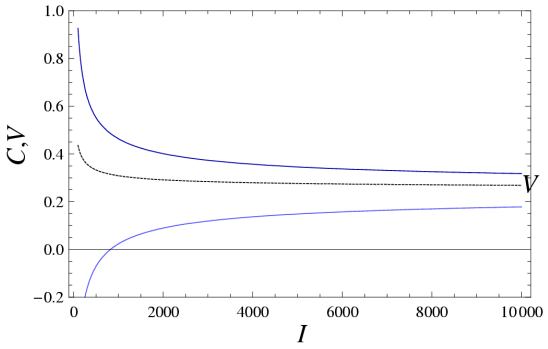

Figure 1: Illustration of upper and lower bounds on the result of cross-validation with respect to the size of sample I. Other constants are: $\eta = 0.01 \Rightarrow 1 - \alpha(\eta) \approx 0.93$, $N = 100$, $n = 3$. With probability $1 - \alpha(\eta)$ or greater, the result C of cross-validation falls between the bounds.

Proof of Theorem 2. We remind: $I' = \frac{n-1}{n} I$, $I'' = \frac{1}{n} I$.

With probability at least $1 - \eta$, the following bound on true risk holds true:

$$R(\omega_{I'}) \leq R'_{\text{emp}}(\omega_{I'}) + \sqrt{\frac{\ln N - \ln \eta}{2I'}}. \quad (19)$$

For the selected function $\omega_{I'}$, fixed from now on, Chernoff inequality is satisfied on the testing set (empirical testing risk) in either of its one-side-versions:

$$R''_{\text{emp}}(\omega_{I'}) - R(\omega_{I'}) \leq \sqrt{\frac{-\ln \eta}{2I''}}, \quad (20)$$

$$R(\omega_{I'}) - R''_{\text{emp}}(\omega_{I'}) \leq \sqrt{\frac{-\ln \eta}{2I''}}, \quad (21)$$

with probability at least $1 - \eta$ each. By joining (19) and (20) we obtain, with probability at least[10] $1 - 2\eta$ the system of inequalities:

$$R''_{\text{emp}}(\omega_{I'}) - \sqrt{\frac{-\ln \eta}{2I''}} \leq R(\omega_{I'}) \leq R'_{\text{emp}}(\omega_{I'}) + \sqrt{\frac{\ln N - \ln \eta}{2I'}}. \quad (22)$$

After n independent folds we obtain, with probability at least $(1 - 2\eta)^n$:

$$\underbrace{\frac{1}{n}\sum_{k=1}^{n} R''_{\text{emp}}(\omega_{I',k})}_{C} \leq \frac{1}{n}\sum_{k=1}^{n} R'_{\text{emp}}(\omega_{I',k}) + \sqrt{\frac{\ln N - \ln \eta}{2I'}} + \sqrt{\frac{-\ln \eta}{2I''}}. \quad (23)$$

[10] The minimum probability must be $1 - 2\eta$ rather than $(1 - \eta)^2$ (probabilistic independence case) due to correlations between inequalities. It can be also viewed as the consequence of Bernoulli's inequality.

To conclude the proof, we need to relate somehow $R'_{\text{emp}}(\omega_{I',k})$ from each fold to $R_{\text{emp}}(\omega_I)$. We need the relation in the direction $R'_{\text{emp}}(\omega_{I',k}) \leq \cdots$, so that we can plug the right-hand-side of it into (23) and keep it true. Intuitively, one might expect that choosing an optimal function on a larger sample leads to a greater empirical risk comparing to a smaller sample, i.e. $R_{\text{emp}}(\omega_I) \geq R'_{\text{emp}}(\omega_{I',k})$, because it is usually easier to fit fewer data points using models of equally rich complexities. But we don't know with what probability that occurs. Contrarily, on may easily find a specific data subset for which $R_{\text{emp}}(\omega_I) \leq R'_{\text{emp}}(\omega_{I',k})$.

Lemma 1. *With probability* 1*, true is the following inequality:*

$$\sum_{i=1}^{l'} Q(\mathbf{z}'_i, \omega_{I'}) \leq \sum_{i=1}^{I} Q(\mathbf{z}_i, \omega_I). \tag{24}$$

On the level of sums of errors, not means, the total error for a larger sample will always surpass the total error for a smaller sample. This gives us $I' R'_{\text{emp}}(\omega_{I'}) \leq I R_{\text{emp}}(\omega_I)$ and further:

$$R'_{\text{emp}}(\omega_{I'}) \leq \frac{n}{n-1} R_{\text{emp}}(\omega_I). \tag{25}$$

Unfortunately it is of no use, because of the coefficient $\frac{n}{n-1}$. Thinking of $C-V$ in the theorem, we need a relation with coefficients 1 at both C and V.

In (Vapnik, 1995b, pp. 124) we find the following helpful assertion:

Lemma 2. *With probability at least* $1-2\eta$*:*

$$\int_{\mathbf{Z}} Q(\mathbf{z}, \omega_I) dF(\mathbf{z}) - \underbrace{\inf_{1 \leq j \leq N} \int_{\mathbf{Z}} Q(\mathbf{z}, \omega_j) dF(\mathbf{z})}_{R(\omega_0)} \leq \sqrt{\frac{\ln N - \ln \eta}{2I}} + \sqrt{\frac{-\ln \eta}{2I}} \tag{26}$$

— the true risk for the selected function ω_I *is not farther from the minimal possible risk for this set of functions than* $\sqrt{\frac{\ln N - \ln \eta}{2I}} + \sqrt{\frac{-\ln \eta}{2I}}$.

Proof of that statement given by Vapnik is based on two inequalities (each with probability at least $1-\eta$), the first is (13) — we repeat it here, and the second is Chernoff inequality for the best function ω_0:

$$R(\omega_I) - R_{\text{emp}}(\omega_I) \leq \sqrt{\frac{\ln N - \ln \eta}{2I}}, \tag{27}$$

$$R_{\text{emp}}(\omega_0) - R(\omega_0) \leq \sqrt{\frac{-\ln \eta}{2I}}. \tag{28}$$

And since, by definition of ω_I, $R_{\text{emp}}(\omega_0) \geq R_{\text{emp}}(\omega_I)$, the (26) follows.

Going back to the cross-validation procedure, we notice that in each single fold the measure R_{emp} corresponds by analogy to the measure R in (26) 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 $\{\mathbf{z}_1, \ldots, \mathbf{z}_I\}$, but still from the perspective of a single cross-validation fold we may view $R_{\text{emp}}(\omega_I)$ as the "target" minimal *probability* of misclassification and $R'_{\text{emp}}(\omega_{I'})$ as the observed relative *frequency* of misclassification — an estimate of that *probability*, remember that we take random subsets $\{\mathbf{z}'_1, \ldots, \mathbf{z}'_{I'}\}$ from the whole set $\{\mathbf{z}_1, \ldots, \mathbf{z}_I\}$.

We write

$$R'_{\text{emp}}(\omega_{I'}) \leq R'_{\text{emp}}(\omega_I) \leq R_{\text{emp}}(\omega_I) + \sqrt{\frac{-\ln \eta}{2I'}}. \tag{29}$$

The first inequality is true with probability 1 by definition of $\omega_{I'}$. The second is a Chernoff inequality, true with probability at least $1-\eta$.

Now, we plug (29) into (23) and obtain with probability $1-(-\sum_{k=1}^{n}\binom{n}{k}(-1)^k(2\eta)^k)-\eta$ or greater:

$$\begin{aligned}
C &\leq \frac{1}{n} n \left(R_{\text{emp}}(\omega_I) + \sqrt{\frac{-\ln \eta}{2I'}} \right) \\
&\quad + \sqrt{\frac{\ln N - \ln \eta}{2I'}} + \sqrt{\frac{-\ln \eta}{2I''}} \\
&= R_{\text{emp}}(\omega_I) + \sqrt{\frac{n}{n-1}} \sqrt{\frac{\ln N - \ln \eta}{2I}} \\
&\quad + \left(\sqrt{n} + \sqrt{\frac{n}{n-1}} \right) \sqrt{\frac{-\ln \eta}{2I}} \\
&= R_{\text{emp}}(\omega_I) + \left(\sqrt{\frac{n}{n-1}} + 1 - 1 \right) \sqrt{\frac{\ln N - \ln \eta}{2I}} \\
&\quad + \left(\sqrt{n} + \sqrt{\frac{n}{n-1}} \right) \sqrt{\frac{-\ln \eta}{2I}} \\
&= V + \left(\sqrt{\frac{n}{n-1}} - 1 \right) \sqrt{\frac{\ln N - \ln \eta}{2I}} \\
&\quad + \left(\sqrt{n} + \sqrt{\frac{n}{n-1}} \right) \sqrt{\frac{-\ln \eta}{2I}}.
\end{aligned}$$

This concludes the proof of theorem 2. □

Proof of Theorem 3. The proof is analogous to the former proof, but we need to write most of the probabilistic inequalities in the different direction.

With probability at least $1-\eta$, the following bound on true risk holds true:

$$R'_{\text{emp}}(\omega_{I'}) \leq R(\omega_{I'}) + \sqrt{\frac{\ln N - \ln \eta}{2I'}}. \tag{30}$$

By joining (30) and (21) we obtain, with probability at least $1-2\eta$ the system of inequalities:

$$R'_{\text{emp}}(\omega_{I'}) - \sqrt{\frac{\ln N - \ln\eta}{2I'}} \le R(\omega_{I'}) \le R''_{\text{emp}}(\omega_{I'}) + \sqrt{\frac{-\ln\eta}{2I''}}. \quad (31)$$

After n independent folds we obtain, with probability at least $(1-2\eta)^n$:

$$\frac{1}{n}\sum_{k=1}^{n} R'_{\text{emp}}(\omega_{I',k}) - \sqrt{\frac{\ln N - \ln\eta}{2I'}} - \sqrt{\frac{-\ln\eta}{2I''}} \le \underbrace{\frac{1}{n}\sum_{k=1}^{n} R''_{\text{emp}}(\omega_{I',k})}_{C}. \quad (32)$$

Again as in the former proof, we need to relate $R'_{\text{emp}}(\omega_{I',k})$ from each fold to $R_{\text{emp}}(\omega_I)$, but now we need the relation to be in the direction $R'_{\text{emp}}(\omega_{I',k}) \ge \cdots$, so that we can plug the right-hand-side of it into (32) and keep it true.

We write

$$R_{\text{emp}}(\omega_I) - \sqrt{\frac{\ln N - \ln\eta}{2I'}} \le R_{\text{emp}}(\omega'_I) - \sqrt{\frac{\ln N - \ln\eta}{2I'}} \le R'_{\text{emp}}(\omega'_I). \quad (33)$$

Reading it from the right-hand-side: the second is a (13)-like inequality but for discrete measures, which is true with probability at least $1-\eta$, and the first inequality is true with probability 1 by definition of ω_I.

Now, we plug (33) into (32) and obtain with probability $1-(-\sum_{k=1}^{n}\binom{n}{k}(-1)^k(2\eta)^k)-\eta$ or greater:

$$\begin{aligned} C &\ge \frac{1}{n} n\left(R_{\text{emp}}(\omega_I) - \sqrt{\frac{\ln N - \ln\eta}{2I'}}\right) - \sqrt{\frac{\ln N - \ln\eta}{2I'}} - \sqrt{\frac{-\ln\eta}{2I''}} \\ &= R_{\text{emp}}(\omega_I) - 2\sqrt{\frac{n}{n-1}}\sqrt{\frac{\ln N - \ln\eta}{2I}} - \sqrt{n}\sqrt{\frac{-\ln\eta}{2I}} \\ &= R_{\text{emp}}(\omega_I) - \left(2\sqrt{\frac{n}{n-1}} - 1 + 1\right)\sqrt{\frac{\ln N - \ln\eta}{2I}} - \sqrt{n}\sqrt{\frac{-\ln\eta}{2I}} \\ &= V - \left(2\sqrt{\frac{n}{n-1}} + 1\right)\sqrt{\frac{\ln N - \ln\eta}{2I}} - \sqrt{n}\sqrt{\frac{-\ln\eta}{2I}}. \end{aligned}$$

This concludes the proof of theorem 3. □

Using theorems 2 and 3 we can also say what sample size I is necessary so that the the difference $C-V$ or $V-C$ is less than or equal to an imposed epsilon ε^*.

Let us denote the right-hand-sides of upper and lower bounds (17) and (18) by ε_{U} and ε_{L} respectively. Now, suppose we want to have $\varepsilon_{\text{U}}(\eta, I, N, n) \le \varepsilon^*_{\text{U}}$. Solving it for I we get

$$I \ge \frac{1}{2\varepsilon^{*\,2}_{\text{U}}}\left(\left(\sqrt{\frac{n}{n-1}} - 1\right)\sqrt{\ln N - \ln\eta} + \left(\sqrt{n} + \sqrt{\frac{n}{n-1}}\right)\sqrt{-\ln\eta}\right)^2 \quad (34)$$

Similarly, if we want to have $\varepsilon_{\text{L}}(\eta, I, N, n) \le \varepsilon^*_{\text{L}}$.

$$I \ge \frac{1}{2\varepsilon^{*\,2}_{\text{L}}}\left(\left(2\sqrt{\frac{n}{n-1}} + 1\right)\sqrt{\ln N - \ln\eta} + \sqrt{n}\sqrt{-\ln\eta}\right)^2 \quad (35)$$

To give an example: say we have a finite set of 100 functions, $N = 100$, we perform a 5-fold cross-validation, $n = 5$, and we choose $\eta = 0.1$ and $\varepsilon^*_{\text{L}} = \varepsilon^*_{\text{U}} = 0.05$. Then it follows that we need a sample of size $I \ge 5832$ so that the cross-validation result is not worse than $V + 0.05$, whereas we need $I \ge 28314$ so that the cross-validation result is not better than $V - 0.05$. And both results are true with probability $1 - \alpha(\eta, n) \approx 0.73$ or greater.

Remark 3. *For the leave-one-out cross-validation, where $n = I$, both the lower and the upper bound loosen to a constant of order $O\left(\sqrt{\frac{-\ln\eta}{2}}\right)$.*

Actually, one can easily see that as we take larger samples $I \to \infty$ and we stick to the leave-one-out cross-validation $n = I$, the coefficient $\sqrt{\frac{n}{n-1}}$ standing at $\sqrt{\frac{\ln N - \ln\eta}{2I}}$ goes to 1, whereas the coefficient $\sqrt{n}$ standing at $\sqrt{\frac{-\ln\eta}{2I}}$ goes to infinity.

One might ask: for what choice of n each bound is the tightest given η, I, N? Treating for a moment n as a continuous variable, we impose the conditions:

$$\frac{\partial\varepsilon_{\text{U}}(\eta, I, N, n)}{\partial n} = 0, \quad \frac{\partial\varepsilon_{\text{L}}(\eta, I, N, n)}{\partial n} = 0,$$

and we get optimal n values:

$$n^*_{\text{U}} = 1 + \left(\frac{\sqrt{\ln N - \ln\eta} + \sqrt{-\ln\eta}}{\sqrt{-\ln\eta}}\right)^{\frac{2}{3}}, \quad (36)$$

$$n^*_{\text{L}} = 1 + \left(\frac{2\sqrt{\ln N - \ln\eta}}{\sqrt{-\ln\eta}}\right)^{\frac{2}{3}}. \quad (37)$$

Note that these values *do not depend* on the sample size I.

2.2 Regression Estimation Learning Task

Now we consider the set of *real-valued* error functions but we still stay with the simplest case when the set has a *finite* number of elements. We give theorems for the *regression estimation* learning task, analogous to the ones for the classification. We skip proofs — the only changes they would require is the assumption of the *bounded* functions, and the use of Hoeffding inequality in the place of Chernoff inequality.

Theorem 4. *Let $\{Q(\mathbf{z},\omega_j)\}_{\omega_j\in\Omega}$, $j=1,2,\ldots,N$, be a finite set of real-valued bounded functions (regression estimation task) of size N, $0\le Q(\mathbf{z},\omega_j)\le B$. Then, for any $\eta>0$, arbitrarily small, there is a small number*

$$\alpha(\eta,n)=\eta-\sum_{k=1}^{n}\binom{n}{k}(-1)^k(2\eta)^k, \qquad (38)$$

and the number

$$\varepsilon(\eta,I,N,n)=\left(2\sqrt{\frac{n}{n-1}}+1\right)B\sqrt{\frac{\ln N-\ln\eta}{2I}} + \left(\sqrt{n}+\sqrt{\frac{n}{n-1}}\right)B\sqrt{\frac{-\ln\eta}{2I}}, \qquad (39)$$

such that:

$$P\left(|V-C|\le\varepsilon(\eta,I,N,n)\right)\ge 1-\alpha(\eta,n). \qquad (40)$$

Theorem 5. *With probability $1-\alpha(\eta,n)$ or greater, the following inequality holds true:*

$$C-V\le\left(\sqrt{\frac{n}{n-1}}-1\right)B\sqrt{\frac{\ln N-\ln\eta}{2I}} + \left(\sqrt{n}+\sqrt{\frac{n}{n-1}}\right)B\sqrt{\frac{-\ln\eta}{2I}}. \qquad (41)$$

Theorem 6. *With probability $1-\alpha(\eta,n)$ or greater, the following inequality holds true:*

$$V-C\le\left(2\sqrt{\frac{n}{n-1}}+1\right)B\sqrt{\frac{\ln N-\ln\eta}{2I}}+B\sqrt{n}\sqrt{\frac{-\ln\eta}{2I}}. \qquad (42)$$

3 THE RELATIONSHIP FOR AN INFINITE SET OF APPROXIMATING FUNCTIONS

The simplest case with a finite number of functions in the set has been generalized by Vapnik (Vapnik, 1995b; Vapnik and Chervonenkis, 1989; Vapnik and Chervonenkis, 1968) onto *infinite* sets with continuum of elements by introducing several notions of the *capacity* of the set of functions: *entropy*, *annealed entropy*, *growth function*, *Vapnik–Chervonenkis dimension*. We remind them in brief.

First of all, Vapnik defines $N^{\Omega}(\mathbf{z}_1,\ldots,\mathbf{z}_I)$ which is the number of all possible *dichotmies* that can be achieved on a fixed sample $\{\mathbf{z}_1,\ldots,\mathbf{z}_I\}$ using functions from $\{Q(\mathbf{z},\omega)\}_{\omega\in\Omega}$. Then, if we relax the sample the following notions of *capacity* can be considered:

1. expected value of $\ln N^{\Omega}$ — *Vapnik-Chervonenkis entropy*:
$$H^{\Omega}(I)=\int_{\mathbf{z}_1\in\mathbf{Z}}\cdots\int_{\mathbf{z}_I\in\mathbf{Z}}\ln N^{\Omega}(\mathbf{z}_1,\ldots,\mathbf{z}_I)\cdot p(\mathbf{z}_1)\cdots p(\mathbf{z}_I)\mathbf{dz}_1\cdots\mathbf{dz}_I,$$
2. ln of expected value of N^{Ω} — *annealed entropy*:
$$H^{\Omega}_{\text{ann}}(I)=\ln\int_{\mathbf{z}_1\in\mathbf{Z}}\cdots\int_{\mathbf{z}_I\in\mathbf{Z}}N^{\Omega}(\mathbf{z}_1,\ldots,\mathbf{z}_I)\cdot p(\mathbf{z}_1)\cdots p(\mathbf{z}_I)\mathbf{dz}_1\cdots\mathbf{dz}_I,$$
3. ln of supremum of N^{Ω} — *growth function*
$$G^{\Omega}(I)=\ln\sup_{\mathbf{z}_1,\ldots,\mathbf{z}_I}N^{\Omega}(\mathbf{z}_1,\ldots,\mathbf{z}_I).$$

It has been proved that:

$$G^{\Omega}(I)=\begin{cases}=\ln 2^I, & \text{dla } I\le h;\\ \le\ln\sum_{k=0}^{h}\binom{I}{k}, & \text{dla } I>h,\end{cases} \qquad (43)$$

where h is the *Vapnik–Chervonenkis dimension*.

It has been shown (Vapnik, 1995b) that

$$H^{\Omega}(I)\overset{\text{(Jensen)}}{\le}H^{\Omega}_{\text{ann}}(I)\le G^{\Omega}(I)\le\ln\sum_{k=0}^{h}\binom{I}{k} \le\ln\left(\frac{eI}{h}\right)^h=h(1+\ln\frac{I}{h}). \qquad (44)$$

And the right-hand-side of (44) can be suitably inserted in the bounds to replace $\ln N$.

We mention that appropriate generalizations from the set of indicator functions (classification) onto sets of real-valued functions (regression estimation) can be found in (Vapnik, 1995b) and are based on the notions of: ε-*finite net*, *set of classifiers* for a fixed real-valued f, *complete set of classifiers* for Ω.

3.1 Classification Learning Task (Infinite Set of Functions)

For shortness, we give only two theorems for bounds on $V-C$ and $C-V$, the bound on $|V-C|$ is their straightforward consequence (analogically as in previous sections).

Theorem 7. *Let $\{Q(\mathbf{z},\omega)\}_{\omega\in\Omega}$ be an infinite set of indicator functions with finite Vapnik–Chervonenkis dimension h. Then, with probability $1-\alpha(\eta,n)$ or greater, the following inequality holds true:*

$$C-V\leq\left(\sqrt{\frac{n}{n-1}}-1\right)\sqrt{\frac{h(1+\frac{2I}{h})-\ln\frac{\eta}{4}}{I}}+\left(\sqrt{n}+\sqrt{\frac{n}{n-1}}\right)\sqrt{\frac{-\ln\eta}{2I}}. \quad (45)$$

Theorem 8. *With probability $1-\alpha(\eta,n)$ or greater, the following inequality holds true:*

$$V-C\leq\left(2\sqrt{\frac{n}{n-1}}+1\right)\sqrt{\frac{h(1+\frac{2I}{h})-\ln\frac{\eta}{4}}{I}}+\sqrt{n}\sqrt{\frac{-\ln\eta}{2I}}. \quad (46)$$

3.2 Regression Estimation Learning Task (Infinite Set of Functions)

Again, for shortness, we give only two theorems for bounds on $V-C$ and $C-V$, the bound on $|V-C|$ is their straightforward consequence (analogically as in previous sections).

Theorem 9. *Let $\{Q(\mathbf{z},\omega)\}_{\omega\in\Omega}$ be an infinite set of real-valued bounded functions, $0\leq Q(\omega,\mathbf{z})\leq B$, with finite Vapnik–Chervonenkis dimension h. Then, with probability $1-\alpha(\eta,n)$ or greater, the following inequality holds true:*

$$C-V\leq\left(\sqrt{\frac{n}{n-1}}-1\right)B\sqrt{\frac{h(1+\frac{2I}{h})-\ln\frac{\eta}{4}}{I}}+\left(\sqrt{n}+\sqrt{\frac{n}{n-1}}\right)\sqrt{\frac{-\ln\eta}{2I}}. \quad (47)$$

Theorem 10. *With probability $1-\alpha(\eta,n)$ or greater, the following inequality holds true:*

$$V-C\leq\left(2\sqrt{\frac{n}{n-1}}+1\right)B\sqrt{\frac{h(1+\frac{2I}{h})-\ln\frac{\eta}{4}}{I}}+\sqrt{n}\sqrt{\frac{-\ln\eta}{2I}}. \quad (48)$$

In practice, bounds (47) and (48) can be significantly tightened by using an estimate $\widehat{B}$ in the place of the most pessimistic B. The estimate $\widehat{B}$ can be found by performing just one fold of cross-validation (instead of n folds) and bounding $\widehat{B}$ by: mean error on the testing set plus a square root implied by the Chernoff inequality:

$$\widehat{B}\leq R''_{\text{emp}}(\omega'_I)+B\sqrt{\frac{-\ln\eta_B}{2I''}}, \quad (49)$$

where η_B is an imposed small probability that (49) is not true. The reasoning behind this remark is that in practice, typical learning algorithms rarely produce functions $f(\mathbf{x},\omega_I)$, in the process of ERM, having high maximal errors. Therefore, we can insert the right-hand-side of (49) into (47) and (48) in the place of B. If this is done, then the minimal overall probability on bounds (47) and (48) should be adjusted to $1-\alpha(\eta,n)-\eta_B$.

4 EXPERIMENTS — BOUNDS CHECKS

Results of three experiments are shown in this section, for the following cases: (1) binary classification, finite set of functions, (2) binary classification, infinite set of functions, (3) regression estimation, infinite set of functions.

4.1 Set of Functions

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

$$f(\mathbf{x},\underbrace{w0,w_1,\ldots,w_K}_{\omega})=\max\{-1,\min\{1,w_0+\sum_{k=1}^{K}w_k\exp\Big(-\frac{\|\mathbf{x}-\mu_k\|^2}{2{\sigma_k}^2}\Big)\}\} \quad (50)$$

where centers μ_k and widths σ_k were generated on random[11] and remained *fixed*. Therefore we have a set of functions linear in parameters $(w_0,w_1,\ldots,w_K)$. As one can see values of f where constrained by ±1. For the classification learning task, the decision boundary was arising as the solution of $f(\mathbf{x},w_0,w_1,\ldots,w_K)=0$. For the regression estimation, we simply looked at the values of $f(\mathbf{x},w_0,w_1,\ldots,w_K)$. Examples of functions from this set are shown in figures 2, 3

4.2 System and Data Sets

As a system $y(\mathbf{x})$ we picked on random a function from a similar class to (50) but *broader*, in the sense

[11] Random intervals: $\mu_k\in[0,1]^2$, $\sigma_k\in[0.02,0.1]$.

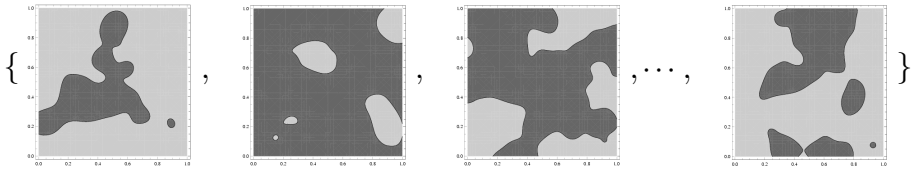

Figure 2: Illustration of the set of functions for classification.

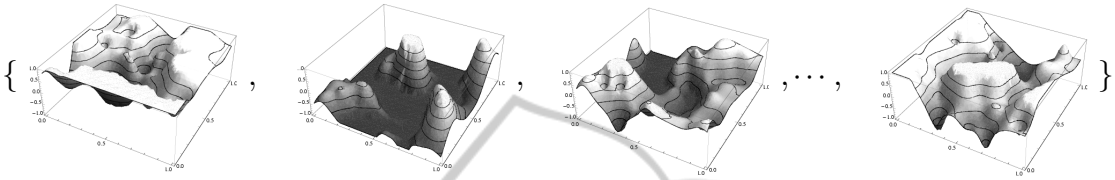

Figure 3: Illustration of the set of functions for regression estimation.

that the number K was greater and the range of randomness on σ_k was larger. Data sets for both classification and regression estimation 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 set $p(\mathbf{x}) = 1$ — uniform distribution on the domain $[0,1]^2$ and $p(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-y(\mathbf{x}))^2}{2\sigma^2})$ — normal noise with $\sigma = 0.1$.

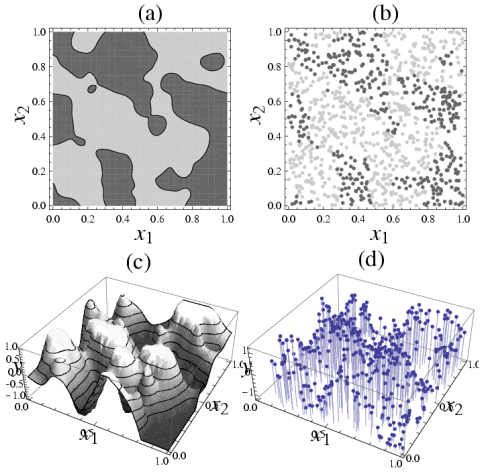

Figure 4: System and data for classification (a, b), regression estimation (c, d).

4.3 Algorithm of the Learning Machine

In the case of *finite* sets of N functions, the learning machine was simply choosing the best functions as $f(\omega_I) = \arg\min_{j=1,2,\ldots,N} R_{\text{emp}}(\omega_j)$ or in cross-validation folds $f(\omega_{I'}) = \arg\min_{j=1,2,\ldots,N} R'_{\text{emp}}(\omega_j)$.

In the case of *infinite* sets with continuum of elements, the learning machine was trained by the least-squares criterion. We remark that obviously other learning approaches can be used in this place e.g. maximum likelihood, SVM criterion (Vapnik, 1995b; Vapnik, 1995a; M. Korzeń and Klęsk, 2008). If we denote the bases $\exp\big(-\frac{\|\mathbf{x}-\mu_k\|^2}{2\sigma_k{}^2}\big)$ by $g_k(\mathbf{x})$ and calculate the matrix of bases at data points

$$G = \begin{pmatrix} 1 & g_1(\mathbf{x}_1) & g_2(\mathbf{x}_1) & \cdots & g_K(\mathbf{x}_1) \\ 1 & g_1(\mathbf{x}_2) & g_2(\mathbf{x}_2) & \cdots & g_K(\mathbf{x}_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & g_1(\mathbf{x}_I) & g_2(\mathbf{x}_I) & \cdots & g_K(\mathbf{x}_I) \end{pmatrix} \quad (51)$$

we can find the optimal vector of w coefficients by the pseudo-inverse operation as follows:

$$(w_0, w_1, \ldots, w_K)^T = (G^T G)^{-1} G^T Y, \quad (52)$$

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

4.4 Experiment Results and Comments

Experiments involved trying out different settings on all relevant constants such as: number of terms in approximating functions (K), number of functions (N) in the case of finite sets or VC dimension (h) in case of infinite sets, sample size (I), number of cross-validation folds (n). For each fixed setting of the constants, an experiment with repetitions was performed, during which we measured the cross-validation outcome C after each repetition. The range of these outcomes was then compared to the interval implied by the theorems we proved.

We show the results in two tables 1 and 2. The first one gives an insight on details of a *single* exemplary experiment: results of its particular folds and repetitions. The second one shows collective results, where each row encapsulates 10 repetitions[12].

[12]It was difficult to allow ourselves for more repeti-

Table 2: Collective results — each row encapsulates 10 repetitions. Tasks: *c.* — classification, *r.e.* — regression estimation. We denote experiments on finite or infinite sets of functions by setting either N or h. For regression estimation we use probabilistic $\widehat{B}$ calculated as $R''_{\text{emp}}(\omega'_I)+B\sqrt{-\ln\eta_B/(2I'')}$. In all experiments $\eta=0.2$, hence for $n=3$ the probability that bounds are true is $1-\alpha(\eta,n)=0.496$ or greater and for $n=5$ it is $1-\alpha(\eta,n)=0.511$ or greater.

no. of exp.	task	K	N	h	I	$R_{\text{emp}}(\omega_I)$	V	n	bounds $[V-\varepsilon_{\text{L}},V+\varepsilon_{\text{U}}]$	observed range of C (10 repetitions)	ratio of C inside bounds
1	c.	50	10	-	10^3	0.412	0.456	3	[0.254, 0.550]	[0.351, 0.445]	1.0
2	c.	200	10	-	10^3	0.345	0.389	3	[0.187, 0.483]	[0.352, 0.385]	1.0
3	c.	200	10	-	10^4	0.369	0.383	3	[0.319, 0.413]	[0.371, 0.383]	1.0
4	c.	200	10	-	10^4	0.396	0.410	5	[0.344, 0.442]	[0.386, 0.401]	1.0
5	c.	50	100	-	10^4	0.408	0.426	3	[0.349, 0.456]	[0.392, 0.418]	1.0
6	c.	200	100	-	10^4	0.336	0.354	3	[0.277, 0.384]	[0.332, 0.338]	1.0
7	c.	50	100	-	10^5	0.401	0.407	3	[0.383, 0.417]	[0.398, 0.403]	1.0
8	c.	50	-	51	10^5	0.181	0.250	3	[0.021, 0.267]	[0.181, 0.184]	1.0
9	c.	200	-	201	10^5	0.035	0.161	3	[−0.25, 0.185]	[0.035, 0.037]	1.0
9	r.e.	50	-	51 ($\widehat{B}=0.193$)	10^4	0.172	0.209	3	[0.078, 0.223]	[0.170, 0.173]	1.0
10	r.e.	50	-	51 ($\widehat{B}=0.194$)	10^4	0.171	0.208	5	[0.085, 0.212]	[0.170, 0.172]	1.0
11	r.e.	200	-	201 ($\widehat{B}=0.020$)	10^5	0.012	0.015	3	[0.006, 0.016]	[0.012, 0.013]	1.0
12	r.e.	200	-	201 ($\widehat{B}=0.020$)	10^5	0.013	0.015	5	[0.007, 0.016]	[0.012, 0.013]	1.0

Table 1: Details (folds, repetitions) of an exemplary experiment no. 1.

no. of experiment	repetition	fold	$R'_{\text{emp}}(\omega_{I'})$	is $\omega_{I'}=\omega_I$?	$R''_{\text{emp}}(\omega_{I''})$
1	1	1	0.397	false	0.444
1	1	2	0.418	true	0.369
1	1	3	0.400	false	0.468
					C = 0.417
1	2	1	0.359	true	0.369
1	2	2	0.374	true	0.339
1	2	3	0.370	true	0.348
					C = 0.352
⋮	⋮	⋮	⋮	⋮	⋮
1	10	1	0.403	true	0.384
1	10	2	0.395	true	0.399
1	10	3	0.394	true	0.399
					C = 0.394

To comment on the results we first remark that before each single experiment (1-12) the whole data set was drawn once from $p(\mathbf{z})$ and remained fixed throughout repetitions. However, in the repetitions due to the non-stratified cross-validation we parted the data set (via permutations) into different training and testing subsets. That is why in the table $R_{emp}(\omega_I)$ and V are constant per experiment, whereas the cross-validation varies within some observed range. In the table 2 we also present the interval $[V-\varepsilon_{\text{L}},V+\varepsilon_{\text{U}}]$ which is implied by the theorems.

Please note that for *all* experiments the observed range for C was contained inside $[V-\varepsilon_{\text{L}},V+\varepsilon_{\text{U}}]$ — an empirical confirmation of theoretical results. Although the bounds are true with probability at least $1-\alpha(\eta,n)$, in this particular experiment they held with frequency one.

In particular one can note in the table that the upper bounds $V+\varepsilon_U$ are closer to actual C outcomes, while lower bounds $V+\varepsilon_L$ are more loose — a fact we already indicated in theoretical sections. Only in the case of experiment no. 9 the lower bound we obtained was trivial. In the results one can also observe the qualitative fact that both intervals tighten with $1/\sqrt{I}$ approximately. Keep in mind that this result stops working for the 'leave-one-out' cross-validation (or a close one) and we experimented on $n=3$ and $n=5$.

5 EXPERIMENTS — SRM

In this section we show results of the *Structural Risk Minimization* approach. We consider a *structure* i.e. a sequence of nested subsets of functions: $S_1\subset S_2\subset\cdots\subset S_K$, where each successive $S_k=\{f(\mathbf{x},\omega)\}_{\omega\in\Omega_k}$ is a set of functions with Vapnik-Chervonenkis dimension h_k, and we have $h_1<h_2<\cdots<h_K$. As the best element of the structure we choose S^* (with VC dimension h^*) for which the bound on generalization V is the smallest.

tions, say 100, due to large amount of results and the time-consumption of each experiment. Yet, the observed ratio 1.0 of C falling inside bounds shows that 10 repetitions was sufficient.

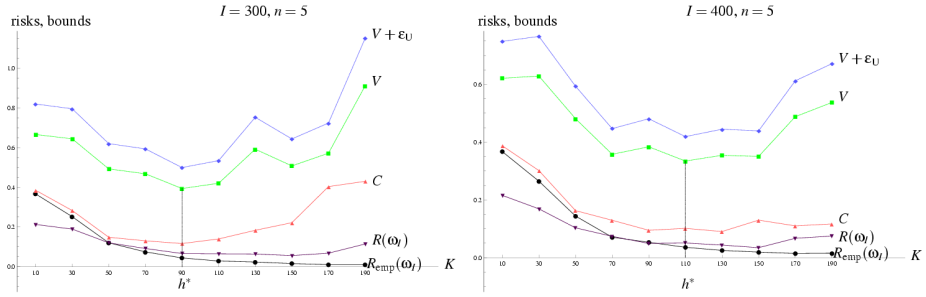

Figure 5: SRM experiments. With $I = 300$, optimum points reached at: $h^* = 91$ (SRM), $h = 91$ (C), $h = 151$ (true risk R). With $I = 400$, optimum points reached at: $h^* = 111$ (SRM), $h = 131$ (C), $h = 151$ (true risk R).

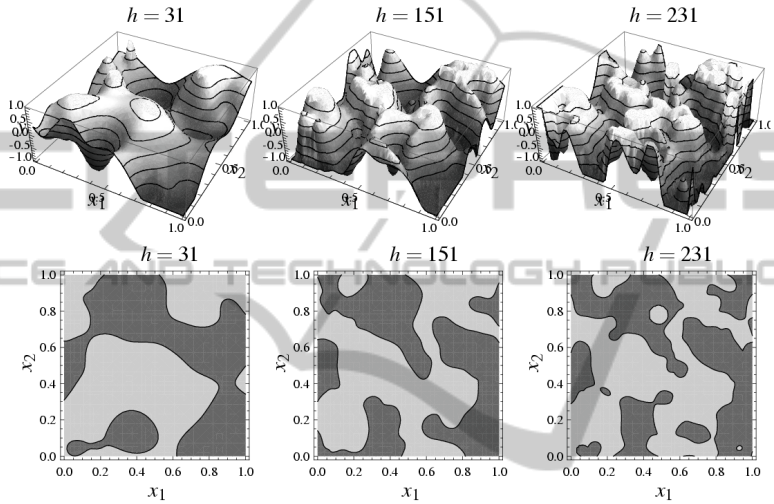

Figure 6: Exemplary models for both regresssion estimation and classfication: under complex ($h = 31$), accurately complex — the best generalization ($h = 151$), over complex ($h = 231$).

Along with observing the bound V, we observe: (1) the cross-validation result C, (2) our bounds on C, (3) the actual *true risk* R calculated as an integral according to its definition (1). We pay particular attention to how the minimum point of SRM at h^* differs from the minimum suggested by the cross-validation and the minimum of true risk (which normally in practice is unknown). We remind that obtaining the result C for each h_k is $O(n)$ times more laborious than obtaining V for each h_k. See fig. 5.

6 SUMMARY

In the paper we take under consideration the probabilistic relationship between two quantities: Vapnik generalization bound V and the result C of an n-fold non-stratified cross-validation. In the literature on the subject of machine learning (and SLT) typically the stated results have a different focus — namely, the relation between the *true risk* (generalization error) and either of the two quantities V, C separately. The perspective we chose was intended to:

- stay in the setting of Structural Risk Minimization approach based on Vapnik bounds,
- *not perform* the cross-validation procedure,
- be able to make probabilistic statements about closeness of SRM results to cross-validation results (if such was perfomed) for given conditions of learning experiment.

Suitable theorems about this relationship are stated and proved. The theorems concern two learning tasks: classification and regression estimation; and also two cases as regards the capacity of the set of approximating functions: finite sets and infinite sets (but with finite Vapnik-Chervonenkis dimension).

As the sample size grows large, both C and V *converge in probability* to the same limit of true risk. The rate of convergence is exponential.

Using the theorems, one can find a threshold size of sample so that the difference $C - V$ or $V - C$ is

smaller than an imposed ε. Obviously, the smaller ε for given experiment conditions, the more frequently one can expect to select the same optimal model complexity via SRM and via cross-validation (again without actually performing it).

For the special case of leave-one-out cross-validation we observe in the consequence of bounds we derived that at most a constant difference of order $O(\sqrt{-\ln\eta/2})$ between C and V can be expected.

Additionally, we showed for what number n of folds, the bounds (lower and upper) on the difference are the tightest. Interestingly, as it turns out these optimal n values *do not* depend on the sample size.

Finally, shown are experiments confirming statistical correctness of the bounds.

ACKNOWLEDGEMENTS

This work has been financed by the Polish Government, Ministry of Science and Higher Education from the sources for science within years 2010–2012. Research project no.: N N516 424938.

REFERENCES

Anthony, M. and Shawe-Taylor, J. (1993). A result of vapnik with applications. *Discrete Applied Mathematics*, 47(3):207–217.

Bartlett, P. (1997). The sample complexity of pattern classification with neural networks: the size of weights is more important then the size of the network. *IEEE Transactions on Information Theory*, 44(2).

Bartlett, P., Kulkarni, S., and Posner, S. (1997). Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 47:1721–1724.

Cherkassky, V. and Mulier, F. (1998). *Learning from data*. John Wiley & Sons, inc.

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

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

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

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

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

Holden, S. (1996a). Cross-validation and the pac learning model. Technical Report RN/96/64, Dept. of CS, University College, London.

Holden, S. (1996b). Pac-like upper bounds for the sample complexity of leave-one-out cross-validation. In *9-th Annual ACM Workshop on Computational Learning Theory*, pages 41–50.

Kearns, M. (1995a). A bound on the error of cross-validation, with consequences for the training-test split. In *Advances in Neural Information Processing Systems 8*. MIT Press.

Kearns, M. (1995b). An experimental and theoretical comparison of model selection methods. In *8-th Annual ACM Workshop on Computational Learning Theory*, pages 21–30.

Kearns, M. and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11:1427–1453.

Kohavi, R. (1995). A study of cross-validation and boostrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Krzyżak, A. et al. (2000). Application of structural risk minimization to multivariate smoothing spline regression estimates. *Bernoulli*, 8(4):475–489.

M. Korzeń, M. and Klęsk, P. (2008). Maximal margin estimation with perceptron-like algorithm. In L. Rutkowski, R. Tadeusiewicz R., L. Z. J. Z., editor, *Lecture Notes in Artificial Intelligence*, pages 597–608. Springer.

Ng, A. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. In *21-st International Conference on* Machine learning, ACM International Conference *Proceeding Series*, volume 69.

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. et al. (1996). A framework for structural risk minimization. *COLT*, pages 68–76.

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

Vapnik, V. (1995b). *Statistical Learning Theory: Inference from Small Samples*. Wiley, New York.

Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data*. Information Science & Statistics. Springer, US.

Vapnik, V. and Chervonenkis, A. (1968). On the uniform convergence of relative frequencies of events to their probabilities. *Dokl. Aka. Nauk*, 181.

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*, 2:217–249.

Weiss, S. and Kulikowski, C. (1991). *Computer Systems That Learn*. Morgan Kaufmann.