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. Zolnierska 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 \( \epsilon \)-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 Muller, 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 practitioners) 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: \( P(|I - p| \geq \epsilon) \leq 2\exp(-2\epsilon^2 I) \).
Hoeffding inequality: \( P(|X - EX| \geq \epsilon) \leq 2\exp(-\frac{\epsilon^2 I}{2I}) \).

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 an in-probability-convergence and its rate is exponential.
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 $(\beta_1, \beta_2)$-error stability\(^2\) rather than $(\beta_1, \beta_2)$-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.

### 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:

  \[
  \{(x_1, y_1), (x_2, y_2), \ldots, (x_I, y_I)\},
  \]

  or more shortly by encapsulating pairs as

  \[
  \{z_1, z_2, \ldots, z_I\},
  \]

  where \(x_i \in \mathbb{R}^d\) 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:

  \[
  \{f(x, \omega)\}_{\omega \in \Omega},
  \]

  where \(\Omega\) is the domain of parameters of this set of functions, so a fixed \(\omega\) can be regarded as an index of a specific function in the set.

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

  \[
  R(\omega) = \int_{x \in \mathbb{X}} \int_{y \in \mathbb{Y}} L(f(x, \omega), y) \left\{ \frac{p(x, y)}{p(x)p(y|x)} \right\} \, dy \, dx,
  \]

  where \(p(x)\) is the distribution density of input points, \(p(y|x)\) is the conditional density of system/phenomenon outputs \(y\) given a fixed \(x\) \(p(x, y) = p(x)p(y|x)\) is the joint distribution density for pairs \((x, y)\). In practice, \(p(x, y)\) is unknown but fixed, and hence we assume the sample \(\{z_1, z_2, \ldots, z_I\}\) 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:

  \[
  L(f(x, \omega), y) = \begin{cases} 0, & \text{for } y = f(x, \omega); \\ 1, & \text{for } y \neq f(x, \omega), \end{cases}
  \]

\(^2\)We say that a learning algorithm has a \((\beta_1, \beta_2)\)-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 \(\beta_1\)-close to each other with probability at least \(1 - \beta_2\). Obviously the smaller both \(\beta_1, \beta_2\) are the more stable the algorithm.

\(^3\)We say that a learning algorithm has a \((\beta_1, \beta_2)\)-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 \(\beta_1\)-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.

\(^4\)Independent, identically distributed.
and the risk functional becomes $R(\omega) = \int_{x \in X} \sum_{y \in Y} L(f(x, \omega), y) p(x, y) dx$. For regression estimation, $L$ is usually chosen as the distance in $L_2$ metric:

$$L(f(x, \omega), y) = (f(x, \omega) - y)^2,$$

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

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

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

- Since only a finite set of samples $\{\mathbf{z}_1, \ldots, \mathbf{z}_l\}$ 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:

$$R_{\text{emp}}(\omega) = \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(x_i, \omega)),$$

and by $\omega_0'$ we denote the index of the function $f(\mathbf{x}, \omega_0')$ such that:

$$R_{\text{emp}}(\omega_0') = \inf_{\omega \in \Omega} R_{\text{emp}}(\omega).$$


- For simplification of notation and further considerations, we introduce replacements:

$$(x, y) = \mathbf{z},$$

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

$$R(\omega) = \int_{x \in X} \int_{y \in Y} Q(\mathbf{z}, \omega) p(\mathbf{z}) d\mathbf{z} p(x, y),$$

and the empirical risk as

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

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 permuting 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 bootstrap (Efron and Tibshirani, 1993). In the classical cross-validation, all $\binom{l}{2}$ pairs of training sets are mutually disjoint (and so are testing sets) and hence folds are independent, 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 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 split of the whole permuted data set $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l\}$. Similarly, empirical risks calculated as follows:

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

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

represent respectively the training error and the testing error, calculated for any function $\omega$.

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

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

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

---

5In 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 $\omega$ and maps them to error values.
all folds, \(k = 1, 2, \ldots, n\), we will write \(\omega_{k}^{\prime}\) 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 \(R_{\text{emp}}^{\prime\prime}\) using functions \(\omega_{k}^{\prime}\):

\[
C = \frac{1}{n} \sum_{k=1}^{n} R_{\text{emp}}^{\prime\prime}(\omega_{k}^{\prime}). \quad \text{(12)}
\]

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 \(A, B\):

\[
P(R_{\text{emp}}^{\prime\prime}(\omega_{k}^{\prime}) = A, R_{\text{emp}}^{\prime\prime}(\omega_{l}^{\prime}) = B) = P(R_{\text{emp}}^{\prime\prime}(\omega_{k}^{\prime}) = A) \cdot P(R_{\text{emp}}^{\prime\prime}(\omega_{l}^{\prime}) = B).
\]

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 learning task and the simplest case of a function. Similarly to Vapnik, let us start with the classification learning task.

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

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

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

**Remark 2.** For the fixed values of \(\eta, N, n\), 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 $\eta \to 0^+$

$$\sqrt{\ln N} - \ln \eta \sim \sqrt{-\ln \eta}.$$  

Therefore, for fixed $\eta, N, n$ there exists a constant, say $D$, such that

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

$$+ \left( \sqrt{n} + \sqrt{\frac{n}{n-1}} \right) \sqrt{-\ln \eta} \leq D \sqrt{-\ln \eta}.$$  

Solving the inequality for $\eta$ we obtain $\eta \leq \exp(-2Ie^2/D^2)$. \(\square\)

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:

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

$$+ \left( \sqrt{n} + \sqrt{\frac{n}{n-1}} \right) \sqrt{-\ln \eta}. \quad (17)$$

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

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

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

\(^8\)We say that $A(I)$ converges in probability to $B$, when for any numbers $\varepsilon > 0$, $\eta > 0$, there exists a threshold size of sample $I(\varepsilon, \eta)$, such that for all $I \geq I(\varepsilon, \eta)$:

$$P\{|A(I) - B| > \varepsilon\} \leq \eta.$$

\(^9\)C and $V$ can be viewed as random variables, due to random realizations of data set $\{z_1, \ldots, z_t\}$ with joint density $p(z)$ (this affects $C$ and $V$) and due to random realizations of subsets in cross-validation folds (this affects $C$). When the data set $\{z_1, \ldots, z_t\}$ is fixed, $V$ is fixed too.

**Proof of Theorem 2.** We remind: $I' = \frac{1}{n} I$, $I'' = \frac{1}{n^2} I$. With probability at least $1 - \eta$, the following bound on true risk holds true:

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

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

$$R_{\text{emp}}(\omega_f) - R(\omega_f) \leq \sqrt{\frac{\ln N - \ln \eta}{2I'}}. \quad (20)$$

$$R(\omega_f) - R_{\text{emp}}(\omega_f) \leq \sqrt{\frac{\ln N - \ln \eta}{2I'}}. \quad (21)$$

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:

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

$$+ \sqrt{\frac{\ln N - \ln \eta}{2I'}}. \quad (22)$$

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

$$\frac{1}{n} \sum_{k=1}^{n} R_{\text{emp}}(\omega_f, k) \leq \frac{1}{n} \sum_{k=1}^{n} R_{\text{emp}}(\omega_f, k) + \sqrt{\frac{\ln N - \ln \eta}{2I'}}$$

$$+ \sqrt{\frac{\ln N - \ln \eta}{2I'}}. \quad (23)$$

\(^{10}\)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 $R'_{\text{emp}}(\omega_{r,k})$ from each fold to $R_{\text{emp}}(\omega_t)$. We need the relation in the direction $R'_{\text{emp}}(\omega_{r,k}) \leq \cdots$, 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 compared to a smaller sample, i.e. $R_{\text{emp}}(\omega_t) \geq R'_{\text{emp}}(\omega_{r,k})$, 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 $R_{\text{emp}}(\omega_t) \leq R'_{\text{emp}}(\omega_{r,k})$.

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

$$\sum_{i=1}^{r} Q(z'_i, \omega_t) \leq \sum_{i=1}^{r} Q(z_i, \omega_t).$$

(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 $\sum_{i=1}^{r} R'_{\text{emp}}(\omega_t) \geq R_{\text{emp}}(\omega_t)$ and further:

$$R'_{\text{emp}}(\omega_t) \leq \frac{n}{n-1} R_{\text{emp}}(\omega_t).$$

(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 $1 - 2\eta$:

$$\int_{Z} Q(z, \omega_{t})dF(z) - \inf_{1 \leq j \leq N} \int_{Z} Q(z, \omega_{j})dF(z) \leq \sqrt{\frac{\ln N - \ln \eta}{2I} + \frac{-\ln \eta}{2I}}.$$  

(26)

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

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 $\omega_0$:

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

(27)

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

(28)

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

Going back to the cross-validation procedure, we notice that in each single fold the measure $R_{\text{emp}}$ corresponds by analogy to the measure $R$ in (26) and the measure $R'_{\text{emp}}$ corresponds by analogy to $R_{\text{emp}}$ therein. Obviously $R$ is defined on an infinite and continuous space $Z = X \times Y$, whereas $R_{\text{emp}}$ is defined on a discrete and finite sample $\{z_1, \ldots, z_I\}$, but still from the perspective of a single cross-validation fold we may view $R_{\text{emp}}(\omega_t)$ as the “target” minimal probability of misclassification and $R'_{\text{emp}}(\omega_t)$ as the observed relative frequency of misclassification — an estimate of that probability, remember that we take random subsets $\{z'_1, \ldots, z'_I\}$ from the whole set $\{z_1, \ldots, z_I\}$.

We write

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

(29)

The first inequality is true with probability 1 by definition of $\omega_t$. The second is a Chernoff inequality, true with probability at least $1 - \eta$.

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

$$C \leq \frac{1}{n} R_{\text{emp}}(\omega_t) + \frac{\sqrt{n} - \ln \eta}{2I} + \sqrt{\frac{-\ln \eta}{2I}}$$

$$= R_{\text{emp}}(\omega_t) + \frac{n}{n-1} \sqrt{\frac{-\ln \eta}{2I}}$$

$$+ \left(\sqrt{n} + \frac{n}{n-1}\right) \sqrt{\frac{\ln N - \ln \eta}{2I}}$$

$$= R_{\text{emp}}(\omega_t) + \left(\sqrt{n} - 1 + 1\right) \sqrt{\frac{\ln N - \ln \eta}{2I}}$$

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

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

$$+ \left(\sqrt{n} + \frac{n}{n-1}\right) \sqrt{\frac{-\ln \eta}{2I}}.$$  

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:

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

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

$$R_{\text{emp}}(\omega_r) - \sqrt{\frac{\ln N - \ln \eta}{2}} \leq R(\omega_r) \leq R_{\text{emp}}''(\omega_r) + \sqrt{\frac{-\ln \eta}{2}}. \quad (31)$$

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

$$\frac{1}{n} \sum_{k=1}^{n} R_{\text{emp}}'(\omega_{r,k}) - \sqrt{\frac{\ln N - \ln \eta}{2}} \leq R_{\text{emp}}'(\omega_{r,k}) \leq \frac{1}{n} \sum_{k=1}^{n} R_{\text{emp}}''(\omega_{r,k}). \quad (32)$$

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

We write

$$R_{\text{emp}}(\omega_r) - \sqrt{\frac{\ln N - \ln \eta}{2}} \leq R_{\text{emp}}(\omega_{r,k}) \leq R_{\text{emp}}''(\omega_{r,k}). \quad (33)$$

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 $\omega_r$.

Now, we plug (33) into (32) and obtain with probability $1 - \left(\frac{n}{N}\right)^3 (1) \leq (1/2 N)^3$ or greater:

$$C \geq \frac{1}{n} \sum_{k=1}^{n} R_{\text{emp}}'(\omega_{r,k}) - \sqrt{\frac{\ln N - \ln \eta}{2}} - \sqrt{\frac{-\ln \eta}{2}}$$

$$= R_{\text{emp}}(\omega_r) - 2 \left(\frac{n}{N}\right)^3 (1 - \frac{1}{2 N}) \sqrt{\ln \frac{-\ln \eta}{2}}$$

$$= R_{\text{emp}}(\omega_r) - 2 \left(\frac{n}{N}\right)^3 (1 - \frac{1}{2 N}) \sqrt{\ln \frac{-\ln \eta}{2}}$$

$$= V - 2 \left(\frac{n}{N}\right)^3 (1 - \frac{1}{2 N}) \sqrt{\ln \frac{-\ln \eta}{2}}$$

This concludes the proof of theorem 3.

Now, suppose we want to have $\varepsilon_U(\eta, I, N, n) \leq \varepsilon_U^*$. Solving it for $I$ we get

$$I \geq \frac{1}{2 \varepsilon_U^*} \left( \frac{n}{\sqrt{N} - 1} \right) \sqrt{\ln N - \ln \eta} \quad \left(\sqrt{\eta} + \sqrt{\frac{n}{N - 1}} \right)^2$$

Similarly, if we want to have $\varepsilon_L(\eta, I, N, n) \leq \varepsilon_L^*$:

$$I \geq \frac{1}{2 \varepsilon_L^*} \left( \frac{n}{\sqrt{N} - 1} \right) \sqrt{\ln N - \ln \eta} + \sqrt{\eta} \sqrt{-\ln \eta}$$

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 $\varepsilon_U^* = \varepsilon_L^* = 0.05$. Then it follows that we need a sample of size $I \geq 5832$ so that the cross-validation result is not worse than $V + 0.05$, whereas we need $I \geq 28314$ 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 $O(\sqrt{-\ln \eta})$.

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 $\sqrt{\frac{\ln N - \ln \eta}{2}}$ goes to 1, whereas the coefficient $\sqrt{n}$ standing at $\sqrt{-\ln \eta}$ goes to infinity.

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

$$\partial \varepsilon_U(\eta, I, N, n) = 0, \quad \partial \varepsilon_L(\eta, I, N, n) = 0,$$

and we get optimal $n$ values:

$$n_U^* = 1 + \left( \frac{\sqrt{\ln N - \ln \eta} + \sqrt{-\ln \eta}}{\sqrt{-\ln \eta}} \right)^2, \quad (36)$$

$$n_L^* = 1 + \left( \frac{2 \sqrt{\ln N - \ln \eta}}{\sqrt{-\ln \eta}} \right)^2. \quad (37)$$

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

\[
\alpha(\eta, n) = \eta - \sum_{k=1}^{n} \binom{n}{k} (-1)^k (2\eta)^k,
\]

and the number

\[
\varepsilon(\eta, l, N, n) = \left(2\frac{\sqrt{n}}{n - 1} - 1\right) B \sqrt{\frac{\ln N - \ln \eta}{2l}} + \left(\sqrt{\eta} + \frac{n}{n - 1}\right) B \sqrt{-\ln \eta \over 2l},
\]

such that:

\[
P\left( |V - C| \leq \varepsilon(\eta, l, N, n) \right) \geq 1 - \alpha(\eta, n).
\]

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

\[
C - V \leq \left(\frac{n}{n - 1} - 1\right) B \sqrt{\frac{\ln N - \ln \eta}{2l}} + \left(\sqrt{\eta} + \frac{n}{n - 1}\right) B \sqrt{-\ln \eta \over 2l}.
\]

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

\[
V - C \leq 2\left(\frac{n}{n - 1} + 1\right) B \sqrt{\frac{\ln N - \ln \eta}{2l}} + 2\sqrt{\eta} \sqrt{-\ln \eta \over 2l}.
\]

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 \( N^\Omega(z_1, \ldots, z_l) \) which is the number of all possible dichotomies that can be achieved on a fixed sample \( \{z_1, \ldots, z_l\} \) using functions from \( \{Q(z, \omega)\}_{\omega \in \Omega}. \) Then, if we relax the sample the following notions of capacity can be considered:

1. expected value of \( \ln N^\Omega \) — Vapnik–Chervonenkis entropy:

\[
H^\Omega(I) = \int_{z_1 \in Z} \cdots \int_{z_l \in Z} \ln N^\Omega(z_1, \ldots, z_l) \cdot p(z_1) \cdots p(z_l) \, dz_1 \cdots dz_l,
\]

2. In of expected value of \( \ln N^\Omega \) — annealed entropy:

\[
H^\Omega_{\alpha}(I) = \int_{z_1 \in Z} \cdots \int_{z_l \in Z} N^\Omega(z_1, \ldots, z_l) \cdot p(z_1) \cdots p(z_l) \, dz_1 \cdots dz_l,
\]

3. In of supremum of \( \ln N^\Omega \) — growth function

\[
G^\Omega(I) = \ln \sup_{z_1, \ldots, z_l} N^\Omega(z_1, \ldots, z_l).
\]

It has been proved that:

\[
G^\Omega(I) = \begin{cases} \ln 2l, & \text{dla } l \leq h; \\ \leq \ln \sum_{k=0}^{h} \binom{l}{k}, & \text{dla } l > h, \end{cases}
\]

where \( h \) is the Vapnik–Chervonenkis dimension.

It has been shown (Vapnik, 1995b) that

\[
H^\Omega(I) \leq H^\Omega_{\alpha}(I) \leq G^\Omega(I) \leq \ln \sum_{k=0}^{h} \binom{l}{k} \leq \ln \left( \frac{e l^{h}}{h} \right) = h(1 + \ln \frac{l}{h}).
\]

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 \( \varepsilon \)-finite net, set of classifiers for a fixed real-valued \( f, \) complete set of classifiers for \( \Omega. \)

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 \( \{Q(z, \omega)\}_{\omega \in \Omega} \) 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:

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

(45)

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

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

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

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

(47)

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

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

(48)

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

\[
\hat{B} \leq R^\text{emp}_{\text{Std}}(\omega_j) + B \sqrt{-\frac{\ln \eta_B}{2P}},
\]  

(49)

where \( \eta_B \) 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 \( f(x, \omega_j) \), 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, \( f : [0, 1]^2 \rightarrow [-1, 1] \), was Gaussian-like:

\[
f(x, w_0, w_1, \ldots, w_K) = \max\{ -1, \min\{ 1, w_0 + \sum_{k=1}^{K} w_k \exp\left( -\frac{\|x - \mu_k\|^2}{2\sigma_k^2} \right) \} \}
\]  

(50)

where centers \( \mu_k \) and widths \( \sigma_k \) 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 \( \pm 1 \). For the classification learning task, the decision boundary was arising as the solution of \( f(x, w_0, w_1, \ldots, w_K) = 0 \). For the regression estimation, we simply looked at the values of \( f(x, w_0, w_1, \ldots, w_K) \). Examples of functions from this set are shown in figures 2, 3

4.2 System and Data Sets

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

\(^{11}\)Random intervals: \( \mu_k \in [0, 1]^2 \), \( \sigma_k \in [0.02, 0.1] \).
that the number $K$ was greater and the range of randomness on $\sigma$ was larger. Data sets for both classification and regression estimation were taken by sampling the system according to the joint probability density $p(x, y) = p(x)p(y|x)$ where we set $p(x) = 1$ — uniform distribution on the domain $[0, 1]^2$ and $p(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-y(x))^2}{2\sigma^2}\right)$ — normal noise with $\sigma = 0.1$.

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

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. Korzen and Klęsk, 2008). If we denote the bases $\exp\left(-\frac{\|x-\mu_k\|^2}{2\sigma_k^2}\right)$ by $g_k(x)$ and calculate the matrix of bases at data points

$$G = \begin{pmatrix} g_1(x_1) & g_2(x_1) & \cdots & g_K(x_1) \\ g_1(x_2) & g_2(x_2) & \cdots & g_K(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_1(x_I) & g_2(x_I) & \cdots & g_K(x_I) \end{pmatrix}$$

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

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

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-
To comment on the results we first remark that before each single experiment (1-12) the whole data set was drawn once from $p(x)$ 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 $R_{emp}(\theta_I)$ and $V$ are constant per experiment, whereas the cross-validation varies within some observed range. In the table 2 we also present the interval $[V - \varepsilon_L, V + \varepsilon_U]$ which is implied by the theorems.

Please note that for all experiments the observed range for $C$ was contained inside $[V - \varepsilon_L, V + \varepsilon_U]$ — 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 $1/\sqrt{T}$ 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 $S_k = \{f(x, \omega_i)\}_{i=1}^{N_k}$ 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.
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 regression estimation and classification: 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 performed) 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 $\varepsilon$. Obviously, the smaller $\varepsilon$ 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 $O(\sqrt{-\ln n/2})$ 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


