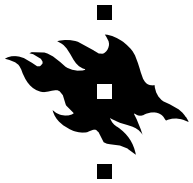
Department of Computer Science
Series of Publications C
Report C-2004-57

Relating the Rademacher and VC Bounds

Matti Kääriäinen



University of Helsinki
Finland

Relating the Rademacher and VC Bounds

Matti Kääriäinen
Department of Computer Science
University of Helsinki, Finland
matti.kaariainen@cs.helsinki.fi

Department of Computer Science, University of Helsinki
Technical report, Series of Publications C, Report C-2004-57
Helsinki, August 2004, i + 12 pages

Abstract

In this technical report we investigate the relationship between generalization error bounds based on VC dimension and Rademacher penalties. We show that a version of the standard VC bound can be recovered from the Rademacher bound, thus providing a direct proof that Rademacher bounds are always at least as good as VC bounds (modulo a small constant factor). The proof highlights in a transparent way the properties of the learning sample that the Rademacher bound takes advantage of but the VC bound overlooks. This clarifies why and when Rademacher penalization yields better results than VC dimension bounds do. As a byproduct we get a new simple proof of the fact that the conditional expectation of Rademacher penalty can be upper bounded by a function of empirical shatter coefficients. Our empirical experiments show that the Rademacher bound can beat VC bounds even when the distribution generating the learning sample is as bad as can be.

Computing Reviews (1998) Categories and Subject Descriptors:
F.2.2 Analysis of Algorithms and Problem Complexity: Nonnumerical Algorithms and Problems
I.2.6 Learning
G.3 Probability and Statistics: Nonparametric Statistics

General Terms:
Theory, Experimentation

Additional Key Words and Phrases:
Pattern Recognition, Learning Theory, Generalization Error Analysis, Rademacher Penalization

1 Introduction

VC bounds are the established method of providing training set bounds for generalization error in statistical learning theory, at least when none of the more specialized bounds apply. However, it is widely known that VC bounds are often too loose to be of any use in practice. As stated by Vapnik [15], finding better constructive bounds for generalization error is thus a very important goal in statistical learning theory. A recent approach towards this goal is Rademacher penalization. By taking the learning sample into account in the complexity penalty, Rademacher bounds [8, 2] may provide tighter estimates of the generalization error when the distribution generating the learning sample is not as bad as can be. Even in case of worst-case distributions, it is known that the Rademacher bound is asymptotically dominated by the VC bound [2]. The Rademacher bound is thus always at least as tight as the VC dimension bound modulo an unknown multiplicative constant, but the quantitative non-asymptotic relationship between the two bounds has been left open.

In this technical report we show by an elementary proof that Rademacher penalty — the complexity penalty term in the Rademacher bound — is always at most $\sqrt{2}$ times as large as the complexity penalty in the tightest known VC bound. This implies that the Rademacher bound is almost as good as the VC bound even in the worst case. The advantage of using a Rademacher bound, on the other hand, may be substantial. This is because the complexity penalties in the VC bounds match the rate of uniform convergence of empirical errors to their expectations for worst-case distributions only, whereas Rademacher penalties track the expected rate of this convergence optimally up to constants for all possible distributions [8]. The advantage of Rademacher bounds for non-worst-case distributions is not only predicted by theory, but also shows up in practice [7, 9]. Our proof of the VC bound demonstrates where exactly this advantage comes from.

Based on the above, it seems that using the VC bound instead of the Rademacher bound can be justified by computational complexity issues only. Evaluating the VC bound requires only a few arithmetical operations provided that an upper bound for the VC dimension is known. On the other hand, the standard method of evaluating Rademacher penalties for a hypothesis class relies on the existence of an *empirical risk minimization (ERM)* algorithm — an algorithm that finds a hypothesis with minimal error rate on the learning sample [8]. Unfortunately, ERM is intractable for many hypothesis classes of practical importance, e.g., in the case of linear classifiers and size restricted decision trees [5].

The computational complexity of ERM has been argued to provide an obstacle to the direct applicability of Rademacher penalization [2]. However, the complexity of evaluating generalization error bounds based on Rademacher penalties may be tractable even if the corresponding ERM task is NP-hard or

even inapproximable. The reason is that all known hardness results for ERM are worst-case results, whereas to evaluate Rademacher bounds it suffices to be able to find the ERM hypothesis in most but not necessarily all cases — and even this is not a necessary condition. The hardness of ERM implies only that the established method for evaluating Rademacher bounds is intractable, but does not rule out the possibility of alternative strategies. Whether such strategies exist seems to be a hard question to decide in either direction. However, as long as the question is open, there is hope that Rademacher bounds can be evaluated efficiently for all important hypothesis classes by some as yet unknown method — even in case P≠NP. If the computational difficulties can be overcome, we claim based on our results that the Rademacher bound has potential to become the standard error bounding technique in situations where VC bounds have traditionally been applied.

2 Background

In this technical report we study generalization error analysis in one variation of *statistical learning theory* [15]. The learning model we use can be described briefly as follows. The *learner* is given a *learning sample* $(x_1, y_1), \ldots, (x_n, y_n)$ of n labeled examples (x_i, y_i), where the *(unlabeled) examples* x_i belong to a set $\mathcal{X}$ and the *labels* y_i are chosen from $\{-1, +1\}$. It is assumed that the labeled examples (x_i, y_i) are independent and identically distributed (i.i.d) realizations of some random variable (X, Y) having an unknown distribution P on $\mathcal{X} \times \{-1, +1\}$. Given a learning sample, the learner has to output a *classifier* $\hat{f} \in F$. Here, F is a *hypothesis class* consisting of an arbitrary set of classifiers, that is, functions mapping $\mathcal{X}$ to $\{-1, +1\}$. The goal of the learner is to select $\hat{f}$ so that it has small *generalization error*, where the generalization error ϵf of an arbitrary classifier f is simply its probability of misclassification:

$$\epsilon(f) = P(f(X) \neq Y).$$

Of course, the generalization error of a classifier cannot be evaluated without knowledge of P. For this reason, one often resorts to *empirical risk minimization* (ERM) [14] and chooses an $\hat{f} \in F$ that has minimal *empirical error*

$$\hat{\epsilon}(f) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{f(x_i) \neq y_i\}}.$$

The theoretical motivation behind the ERM principle is that if

$$\sup_{f \in F} (\epsilon(f) - \hat{\epsilon}(f)) \tag{1}$$

can with high probability be guaranteed to be small, then the generalization error of a classifier with minimal empirical error will with high probability be

close to minimal, too. Upper bounds for (1) — also referred to as complexity penalties in the sequel — can be directly used in providing *generalization error bounds*, that is, upper bounds for $\epsilon(f)$, since

$$\epsilon(f) \leq \hat{\epsilon}(f) + \sup_{f \in F}(\epsilon(f) - \hat{\epsilon}(f)).$$

The focus in the rest of this technical report is on analyzing the relationship between two generalization error bounds of this kind, namely a *VC bound* and a *Rademacher bound*. In order to present these bounds, we first need to define a few underlying concepts.

Definition 1 *Let F be a class of classifiers and $S = \{x_1, \ldots, x_n\}$ a sample of n learning examples. Let $F|S$ denote the set $\{f|S \mid f \in F\}$ of restrictions of functions of F to S.*

- *The* empirical shatter coefficient *of F on S is $N(F, S) = |\{F|S\}|$, and the* shatter coefficient *of F on samples of size n is $N(F, n) = \max\{N(F, S) \mid S \subset \mathcal{X}, |S| \leq n\}$.*
- *The* VC dimension *of F is $VCdim(F) = \max\{n \in \mathbb{N} \mid N(F, n) = 2^n\}$. The* empirical VC dimension *of F on a sample S is the VC dimension of $F|S$, that is, $VCdim(F, S) = VCdim(F|S)$.*

Put in words, $N(F, S)$ is the number of different ways F can classify a sample S, while $N(F, n)$ is its largest value over all samples of size n. Thus, these quantities obviously have something to do with the intuitive concept of complexity of F (on S). VC dimension can be related to the complexity of F by the following lemma [13]:

Lemma 1 (Sauer's lemma) *If F is a class of classifiers with $VCdim(F) \leq d$, then*

$$N(F, n) \leq \sum_{i=1}^{n} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d.$$

For the empirical quantities, if $VCdim(F, S) \leq d = d(S)$, then the above inequalities hold with $N(F, n)$ replaced by $N(F, S)$.

This lemma will be needed in Section 3 when rederiving the VC bound.

Equipped with these definitions, we are ready to state a VC bound for generalization error. This bound due to Devroye [3] is an optimized version of Vapnik's original bound [14]. In particular, the constants in the complexity penalty that determine its rate of convergence w.r.t. n are optimal [4]. The complexity penalty in Vapnik's original bound is essentially the one in Devroye's bound multiplied by 2.

Theorem 1 *Let F be a class of $\{-1,+1\}$-valued functions with VC dimension at most d. With probability at least $1-\delta$, it is true for all $f \in F$ that*

$$\epsilon(f) \leq \hat{\epsilon}(f) + \sqrt{\frac{d(\ln(n^2/d)+1)+\ln(4/\delta)+8}{2n}}.$$

Bounds similar to the above can also be stated in terms of empirical VC dimension, although the constants will be slightly inferior [1]. One can also obtain potentially tighter bounds based on empirical shatter coefficients with similar proof techniques. The bounds based on these empirical quantities are always at least as tight as the VC bound, but because there are no efficient methods for evaluating $N(F,S)$ or $\mathrm{VCdim}(F,S)$ exactly, the bounds have not been widely used or even tested in practice. Data-dependent upper bounds for $\mathrm{VCdim}(F,S)$ and error bounds based on them have received some attention, though [1].

As in the case of VC bounds, there are many slightly different versions of Rademacher penalties and bounds based on them (see, e.g., [2, 8, 1]). We will use a definition that most closely resembles the one given by Bartlett and Mendelson[1] [2].

Definition 2 *Let $\sigma = (\sigma_1, \ldots, \sigma_n)$ be a vector of symmetrical $\{-1,+1\}$-valued random variables that are independent of each other and of the learning sample. The* Rademacher penalty *of a class of classifiers F is*

$$R_n(F) = \frac{1}{n} \sup_{f \in F} \sum_{i=1}^{n} \sigma_i f(X_i).$$

Note that $R_n(F)$ is a random variable depending on the randomly chosen objects X_i and on the random signs σ_i but not on the labels Y_i. Given a realization (S,σ) of the learning sample and the random signs, the supremum in the definition is achieved by the function $f \in F$ whose restriction $f|S$ when viewed as a point in $\{-1,+1\}^n$ has minimal Hamming distance to σ. Equivalently, this f is the ERM hypothesis for the learning sample (S,σ). Thus, the computational problem of evaluating $R_n(F)$ at a point (S,σ) resembles closely ERM, finding the nearest neighbor of a query point in a Hamming cube, finding the closest codeword in decoding an error correcting code, and probably other well-studied computational problems, too.

The following Rademacher bound is a corollary of the generalization error bound presented in [2, Theorem 5 (b)] and can be derived from it by a simple application of McDiarmid's concentration inequality [11].

[1] The difference between our definition and the one given in [2] is that we omit taking the absolute value inside the supremum in order to make the relation between evaluating $R_n(F)$ and ERM more transparent. The change in the definition makes a difference only if F is not closed under complementation, that is, multiplication by -1. Even if F is not closed under complementation, the proof of the generalization error bound in [2] goes through with our definition of $R_n(F)$, too.

Theorem 2 *With probability at least $1-\delta$ over the choice of the learning sample and the random signs $\sigma_1, \ldots, \sigma_n$, it holds for all $f \in F$ that*

$$\epsilon(f) \leq \hat{\epsilon}(f) + R_n(F) + \frac{3}{\sqrt{2}}\sqrt{\frac{\ln 2/\delta}{n}}.$$

The same holds also if $R_n(F)$ is replaced with $\mathbb{E}[R_n(F)|X_1, \ldots, X_n]$.

Note that this bound can be evaluated efficiently based on the learning sample alone if one has a method for efficiently computing $R_n(F)$.

3 Deriving the VC Bound from the Rademacher Bound

The bounds of Theorems 1 and 2 are intimately related because of the similarity of the techniques used in their proofs, most notably the technique of symmetrization by a ghost sample [2]. For a derivation of a VC bound (with non-optimal constants) that directly uses Rademacher penalties, see [10]. However, the understanding concerning the quantitative relationship between VC and Rademacher bounds is still far from complete. To our knowledge, the following Theorem presented in [2] without a proof is the state of the art in this field. A partial version of the quite involved proof can be found in [12], but a full proof has not been published as it would require several pages to present.

Theorem 3 *Let F be a class of classifiers. For all learning samples S it is true that*

$$\mathbb{E}[R_n(F)|S] = O\left(\sqrt{VCdim(F,S)/n}\right)$$

and

$$\mathbb{E}[R_n(F)|S] = O\left(\sqrt{\log N(F,S)/n}\right).$$

Because it is always the case that $\mathrm{VCdim}(F,S) \leq \mathrm{VCdim}(F)$ it follows that the complexity penalties in the Rademacher bounds of Theorem 2 are always at most as large as the penalty in the VC bound times a constant independent of F. However, Theorem 3 does not specify the size of the constant, and the proof sketches of the theorem give reason to believe that the upper bound for the constant obtainable from them is large.

Our rederivation of the VC bound will yield as a side product new bounds for $\mathbb{E}[R_n(F)|S]$ and $R_n(F)$ in terms of VC dimension. The proofs for the bounds are elementary and the bounds contain no hidden constants like the bounds of Theorem 3. Thus, they can be used directly to derive a VC bound for generalization error with constants only slightly inferior to those in Theorem 1.

In the proofs we will work in the Hamming cube $\{-1,+1\}^n$. The key idea is to interpret $F|S$ as a subset of this hyper-cube thus transforming the evaluation of $R_n(F)$ into a purely combinatorial question concerning the *normalized Hamming distance* of a random point $\sigma \in \{-1,+1\}^n$ to the set $F|S$. This distance $d_H(x,y)$ between vectors $x, y \in \{-1,+1\}^n$ is defined by $d_H(x,y) = |\{i \mid x_i \neq y_i\}|/n$. The volume $\mathrm{Vol}(C)$ of a set $C \subset \{-1,+1\}^n$ is given by $\mathrm{Vol}(C) = |C|/2^n$; thus, the volume of a set is its probability under the uniform distribution.

Lemma 2 *Let F be any hypothesis class, let $S \subset \mathcal{X}$ be a learning sample of size n, and let $d = VCdim(F,S)$. Then*

$$\mathbb{E}[R_n(F)|S] \leq 2\sqrt{\frac{\ln N(F,S)}{n}} + \frac{1}{N(F,S)} \leq 2\sqrt{\frac{d(\ln(n/d)+1)}{n}} + \frac{1}{N(F,S)}.$$

Proof. Consider $F|S$ as a subset of the Hamming cube $\{-1,+1\}^S$ which we will equate with $\{-1,+1\}^n$. Place a d_H-ball of radius r at each of the points of $F|S$ and denote the union of these balls by U_r.

We will find an r such that $\mathrm{Vol}(U_r) \leq 1/N(F,S)$. Let $\mathrm{Vol}(B(r))$ denote the volume of a single d_H-ball of radius r. By Hoeffding's inequality [6] $\mathrm{Vol}(B(r)) \leq e^{-2(1/2-r)^2 n}$. Since U_r is the union of $|F|S| = N(F,S)$ such balls, we have $\mathrm{Vol}(U_r) \leq N(F,S) \cdot \mathrm{Vol}(B(r))$. Note that this inequality is tight only if all the points in $F|S$ are so far apart that the balls centered at them do not intersect one another.

By the above approximations, we have $\mathrm{Vol}(U_r) \leq N(F,S)e^{-2(1/2-r)^2 n}$. This upper bound is $1/N(F,S)$ provided that

$$r = 1/2 - \sqrt{\frac{\ln N(F,S)}{n}}.$$

Take now a random $\sigma \in \{-1,+1\}^n$ and consider $R_n(F)(S,\sigma)$, that is, $R_n(F)$ evaluated at the point (S,σ). It is easy to see that if $\sigma \notin U_r$, then $R_n(F)(S,\sigma) \leq 1-2r$. Even if $\sigma \in U_r$, $R_n(F)$ is at most 1, so we have

$$\begin{aligned}\mathbb{E}[R_n(F)|S] = \\ \mathbb{P}(\sigma \notin U_r)\mathbb{E}[R_n(F)|S, \sigma \notin U_r] + \mathbb{P}(\sigma \in U_r)\mathbb{E}[R_n(F)|S, \sigma \in U_r] \\ \leq (1-2r) + \frac{1}{N(F,S)} = \sqrt{\frac{\ln N(F,S)}{n}} + \frac{1}{N(F,S)}.\end{aligned}$$

For the second inequality of the lemma, it suffices to note that by Sauer's lemma $N(F,S) \leq (en/d)^d$. □

In the interesting case $N(F,S) \geq n$ the first part of this result improves on the corresponding inequality of Theorem 3 by providing an explicit and small

constant factor. Our result concerning empirical VC dimension is inferior to that of Theorem 3 by a factor of $\sqrt{\ln n}$. It seems that this factor is unavoidable using our proof technique.

Lemma 3 *Let F have finite VC dimension of at most d. With probability at least $1-\delta$ over the choice of the learning sample and the random signs, it is true that*

$$R_n(F) \leq \sqrt{2}\sqrt{\frac{d(\ln(n/d)+1)+\ln(1/\delta)}{n}}.$$

Proof. The idea of the proof is very similar to the proof of Lemma 2. However, this time the learning sample S of size n will be random. Given an arbitrary S, let $U_{S,r}$ be the union of $|F|S|$ balls of radius r centered at the points of $F|S$. We will choose r independently of S so that $\mathrm{Vol}(U_{S,r}) \leq \delta$ for each set S. This can be done as follows. As in the proof of Lemma 2, $\mathrm{Vol}(U_{S,r}) \leq |F|S| \cdot \mathrm{Vol}(B(r))$, where equality holds only if none of the balls constituting $U_{S,r}$ intersect. Since $\mathrm{VCdim}(F) \leq d$, Sauer's lemma implies $|F|S| \leq (en/d)^d$ for each S. Using this and Hoefding's inequality as in the proof of Lemma 2, we get

$$\mathrm{Vol}(U_{S,r}) \leq \left(\frac{en}{d}\right)^d e^{-2(1/2-r)^2 n} = \delta$$

provided that $r = 1/2 - \sqrt{\frac{d(\ln(n/d)+1)+\ln(1/\delta)}{2n}}$.

Let us combine the sets $U_{S,r}$ into a single set

$$A = A_r = \{(S,\sigma) \mid S \subset \mathcal{X}, \sigma \in U_{S,r}\} \subset \mathcal{X} \times \{-1,+1\}^n.$$

The probability of A can now be estimated as

$$\mathbb{P}(A) = \mathbb{E}[1_A] = \mathbb{E}[\mathbb{E}[1_A|S]] = \mathbb{E}[\mathbb{P}(U_S)] \leq \mathbb{E}[\delta] = \delta.$$

Again, if $(S,\sigma) \notin A$, then $R_n(F)(S,\sigma) \leq 1-2r$, from which the claim follows by substituting r. □

It should be noted that the upper bounds obtained in the previous two lemmas may not be tight even for worst-case distributions. This is because of the slack in Sauer's lemma and Hoeffding's inequality and other technicalities. All these hindrances will be circumvented if the Rademacher penalty is evaluated directly. This may result in significant savings as indicated in our experiments.

Equipped with the lemmas, we can now prove the following:

Theorem 4 *Let F have finite VC dimension of at most d. Then with probability at least $1-\delta$ over the choice of the learning sample, it holds for all $f \in F$ that*

$$\epsilon(f) \leq \hat{\epsilon}(f) + \sqrt{2}\sqrt{\frac{d(\ln(n/d)+1)+\ln(2/\delta)}{n}} + \frac{3}{\sqrt{2}}\sqrt{\frac{\ln(4/\delta)}{n}}. \qquad (2)$$

Proof. By Theorem 2 we have with probability at least $1 - \delta/2$ for all $f \in F$ that

$$\epsilon(f) \leq \hat{\epsilon}(f) + R_n(F) + \frac{3}{\sqrt{2}}\sqrt{\frac{\ln 4/\delta}{n}}. \tag{3}$$

By the previous lemma, with probability at least $1 - \delta/2$,

$$R_n(F) \leq \sqrt{2}\sqrt{\frac{d(\ln(n/d) + 1) + \ln(2/\delta)}{n}}.$$

Combine these and the claim follows. □

Let us compare the bound derived above to existing bounds based on VC dimension. If the VC dimension of F is large, the latter square root term in (2) will be negligible as compared to the term involving the VC dimension. In such cases, the complexity penalty in 2 is essentially $\sqrt{2}$ times the corresponding term in the bound of Theorem 1, and $1/\sqrt{2}$ times the penalty in the original VC dimension bound proved by Vapnik.

In conclusion, the new bound is looser but still quite comparable to the best known VC dimension bound. The bound alone is not of much interest as it is worse than the best existing bound. However, the new derivation makes the relationship between Rademacher and VC bounds more transparent. As argued in the next section, it shows where the potential advantage of using Rademacher penalization comes from.

4 The Advantages of Rademacher Bounds over VC Bounds

The proofs in the previous section highlight two sources of slackness that loosen the VC bound but not the Rademacher bound. Together they cover the slack due to the use of the union bound for probabilities in the standard proof of Vapnik's VC dimension bound [15].

1. In the VC bounds, it is first assumed that equality holds in the inequality $N(F, S) \leq N(F, n)$. Then, an upper bound for $N(F, S)$ is derived by upper bounding $N(F, n)$ by a combination of an upper bound d for $\mathrm{VCdim}(F)$ and an application of Sauer's lemma. This results in the bound $N(F, S) \leq N(F, n) \leq (en/d)^d$. Besides the slackness inherent in Sauer's lemma, extra slack is introduced if the learning sample S is not worst-case and thus $N(F, S) \ll N(F, n)$. Of course, the situation is still worse if the exact VC dimension of F is unknown and it has to be approximated from above — a case quite common in practice.

2. It is assumed that none of the balls centered at the projected classifiers in $F|S$ intersect. This means that the actual structure of the set $F|S$ as a subset of $\{-1,+1\}^n$ is not taken into account. As a worst case approximation, the projected classifiers are assumed to be spread far apart in $\{-1,+1\}^n$. Thus, the possible high correlation among classifiers within subsets of $F|S$ is neglected, even though at least some correlation is usually present. It can be shown that ignoring the correlations can result in a loss of at least a factor $\sqrt{\ln n}$ in the penalty. Further slack is introduced by using Hoeffding's inequality in bounding the volumes of individual balls.

Thus, the proofs in the previous section show that using the Rademacher penalty $R_n(F)$ or its conditional expectation $\mathbb{E}[R_n(F)|S]$ directly instead of VC dimension bounds is advantageous because the Rademacher bound is able to take both the exact size and structure of $F|S$ into account. Using a VC bound means essentially neglecting the information inherent in S and also the properties of F not captured by VC dimension. To our knowledge, the fact that $R_n(F)$ automatically captures both these properties neglected by the VC dimension bound has not been reported before.

The VC dimension bound can be viewed as an approximation for the Rademacher bound. Based only on the VC dimension and the sample size, it provides an upper bound for $R_n(F)$. As a lot of information is lost, the upper bound is necessarily sometimes loose. However, provided that (an upper bound for) the VC dimension of F is known, the bound is trivial to evaluate and thus all the computational problems of evaluating Rademacher bounds are circumvented. The question arises, then, whether one could derive an easy-to-evaluate upper bound for $R_n(F)$ by more clever means than by simply ignoring all information present in the learning sample. There might be computationally efficient ways of providing tight upper bounds for Rademacher bounds even in cases where evaluating the Rademacher bound exactly is intractable. The margin bounds as presented in [1] do this indirectly through bounding the empirical VC dimension of a class of linear classifiers by a function of the margin. Whether something could be gained by a more direct attack is a question for future research.

5 Empirical Experiments

In Section 3 we showed that the Rademacher bound is never much worse than the VC bound even in worst-case situations. In this section, we present experiments that show that the Rademacher bounds can actually be better than VC bounds in such situations, too. This indicates that the performance guarantee given by Theorem 4 may be sometimes overly pessimistic.

Our experimental setting is as follows. As the hypothesis class, we use the

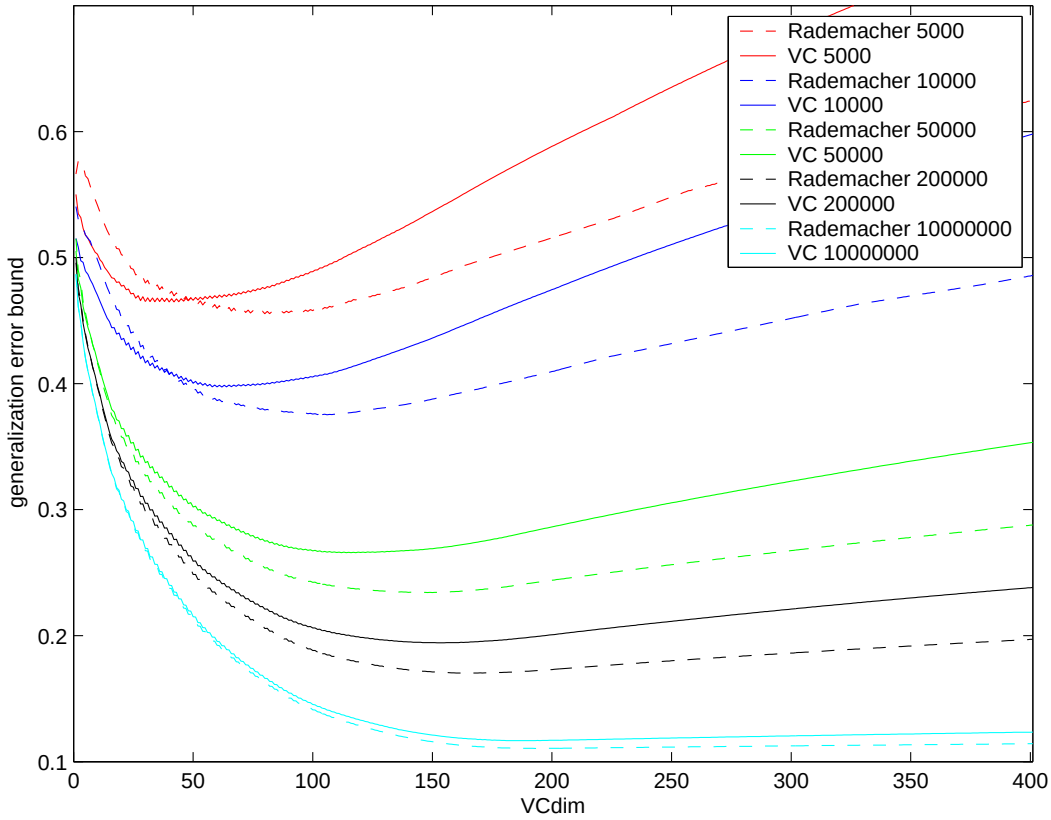


Figure 1: Generalization error bounds for varying sizes of the learning sample. The confidence parameter δ is in all cases 0.01.

class of interval classifiers with at most k splits on the unit interval. It is easy to see that the VC dimension of this class of hypotheses is $k+1$. Also, ERM for this hypothesis class can be easily performed in time $O(nk)$ by dynamic programming. For this reason the same hypothesis class was used by Lozano in his experiments on using Rademacher penalization for model selection [9].

We chose the interval learning problem because with this class of hypotheses, the set $F|S$ is essentially independent of the learning sample S provided that S contains no duplicates. Thus, if we choose the distribution of X to be continuous, this learning task allows us to test whether the Rademacher bound can beat the VC bound when the learning sample provides no extra information on $F|S$.

For a summary of the results of the experiments, see Figure 1. In all experiments, the distribution of X was uniform on $[0,1]$ and the labels were assigned by an interval classifier with 200 splits at random points. The labels thus obtained were corrupted by adding 10% class noise. From the results it is clear that the Rademacher bound is better than the VC bound when n is large or the hypothesis class is complex, i.e., has high VC dimension. If used for model selection, the Rademacher bound suggests choosing a more complex

model than the VC bound does. Thus, even the relatively small difference in the bounds may have significance in practice. Of course, the advantage of the Rademacher bound becomes the larger the easier the distribution is. Thus, if the distribution of X was, e.g., non-continuous and the samples S contained lots of duplicates, the penalty term in the Rademacher bound would decrease significantly whereas the one in the VC bound would remain the same.

Based on the results of the experiments we may conclude that Rademacher penalization can be advantageous even when the distribution generating the learning data is worst-case. As the learning sample provides no extra information for the Rademacher bound on this task, the advantage of the Rademacher bound has to come from two factors. First, as explained before, using Rademacher penalties directly circumvents many approximations necessary in the proof of the VC bound, e.g., the use of Sauer's lemma and Hoeffding's inequality. Second, the Rademacher bound takes into account the correlation between hypotheses, that is, vectors in $F|S$. Even though for this particular F the structure is the same worst-case one for all S (with no duplicates), there is some structure and it can be taken advantage of. We believe that the fact that the VC bound neglects this structure is the prime cause it provides worse bounds on this learning task (and probably on other tasks, too).

References

[1] Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. In *Proceedings of the 13th Annual Conference on Computational Learning Theory*, pages 286–297, San Francisco, 2000. Morgan Kaufmann.

[2] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

[3] Luc Devroye. Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, 12:72–79, 1982.

[4] Luc Devroye, Lázló Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer, 1996.

[5] Michelangelo Grigni, Vincent Mirelli, and Christos H. Papadimitriou. On the difficulty of designing good classifiers. *SIAM J. Comput.*, 30(1):318–323, 2000.

[6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of American Mathematical Association*, 58:13–30, 1963.

[7] Matti Kääriäinen and Tapio Elomaa. Rademacher penalization over decision tree prunings. In *Proc. 14th European Conference on Machine Learning*, pages 193–204, 2003.

[8] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theor.*, 47(?):1902–1914, 2001.

[9] Fernando Lozano. Model selection using Rademacher penalization. In *Proc. 2nd ICSC Symposium on Neural Networks*. NAISO Academic Press, 2000.

[10] Gabor Lugosi. Pattern recognition and learning theory. In *Principles of Nonparametric Learning*. Springer-Verlag, 2002.

[11] Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, 1989.

[12] Shahar Mendelson. ℓ-norm and its application to learning theory. *Positivity*, 5:177–191, 2001.

[13] Nicolas Sauer. On the densities of families of sets. *Journal of Combinatorial Theory - Series A*, 13:145–147, 1972.

[14] Vladimir N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer, 1982.

[15] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.