Journal of Machine Learning Research 3 (2002) 73–98 Submitted 5/02; Published 7/02

Data-dependent margin-based generalization bounds for classification

András Antos ANTOS@MAST.QUEENSU.CA
Department of Mathematics and Statistics,
Queen's University,
Kingston, Ontario, Canada K7L 3N6

Balázs Kégl KEGL@IRO.UMONTREAL.CA
Department of Computer Science and Operational Research,
University of Montreal,
C.P. 6128 Succ. Centre-Ville, Canada, H3C 3J7

Tamás Linder LINDER@MAST.QUEENSU.CA
Department of Mathematics and Statistics,
Queen's University,
Kingston, Ontario, Canada K7L 3N6

Gábor Lugosi LUGOSI@UPF.ES
Department of Economics,
Pompeu Fabra University,
Ramon Trias Fargas 25-27, 08005 Barcelona, Spain

Editor: Peter Bartlett

Abstract

We derive new margin-based inequalities for the probability of error of classifiers. The main feature of these bounds is that they can be calculated using the training data and therefore may be effectively used for model selection purposes. In particular, the bounds involve empirical complexities measured on the training data (such as the empirical fat-shattering dimension) as opposed to their worst-case counterparts traditionally used in such analyses. Also, our bounds appear to be sharper and more general than recent results involving empirical complexity measures. In addition, we develop an alternative data-based bound for the generalization error of classes of convex combinations of classifiers involving an empirical complexity measure that is easier to compute than the empirical covering number or fat-shattering dimension. We also show examples of efficient computation of the new bounds.

Keywords: classification, margin-based bounds, error estimation, fat-shattering dimension

1. Introduction

A large body of recent research on classification focuses on developing upper bounds on the probability of misclassification of a classifier which may be computed using the same data that were used to design the classifier. An interesting family of such bounds is based on "margins", that is, on the confidence a classifier assigns to each well-classified data point. It was already pointed out by Vapnik and Chervonenkis (1974) that usual error bounds based on the VC dimension may be improved significantly in the case of linear classifiers that classify the data well with a large margin. This idea, in turn, has lead to the development of Support Vector Machines (see Vapnik, 1998). Similar, but more general, bounds have been derived based on the notion of the *fat shattering* dimension (see Anthony and Bartlett, 1999, for a survey). The main purpose of this paper is to obtain improved bounds which depend on a data-dependent version of the fat shattering dimension. The new bounds may improve the obtained estimates significantly for many distributions appearing in practice.

Suppose the feature space $\mathcal{X}$ is a measurable set and the observation X and its label Y form a pair (X, Y) of random variables taking values in $\mathcal{X} \times \{0, 1\}$. Let $\mathcal{F}$ be a class of real measurable functions on $\mathcal{X}$. For $f \in \mathcal{F}$, let $L(f)$ denote the probability of error of the prediction rule obtained by thresholding $f(X)$ at $1/2$, that is,

$$L(f) = \mathbf{P}\big\{\operatorname{sgn}(f(X) - 1/2) \neq Y\big\}$$

where

$$\operatorname{sgn}(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{if } t < 0. \end{cases}$$

The margin of f on $(x, y) \in \mathcal{X} \times \{0, 1\}$ is defined by

$$\operatorname{margin}(f(x), y) = \begin{cases} f(x) - 1/2 & \text{if } y = 1 \\ 1/2 - f(x) & \text{if } y = 0. \end{cases}$$

Let the data $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ consist of independent and identically distributed (i.i.d.) copies of (X, Y). For $f \in \mathcal{F}$ and $\gamma > 0$, define the sample error of f on D_n with respect to γ as

$$\widehat{L}_n^\gamma(f) = \frac{1}{n} \sum_{i=1}^{n} I_{\{\operatorname{margin}(f(X_i), Y_i) < \gamma\}}$$

where I_A denotes the indicator of an event A.

It is well known that in many cases $L(f)$ may be upper bounded by the margin error $\widehat{L}_n^\gamma(f)$ plus a quantity that typically decreases with increasing γ, see, for example, Bartlett (1998), Anthony and Bartlett (1999), Shawe-Taylor et al. (1998). In particular, covering numbers and the fat-shattering dimension of $\mathcal{F}$ at scale γ have been used to obtain useful bounds on the probability of error of classifiers. In this paper we develop improved, data-dependent bounds, and show that the empirical version of the fat-shattering dimension may also be used to bound the probability of error.

Our bounds are closely related to results of Shawe-Taylor and Williamson (1999) who obtained generalization bounds in terms of the margin error $\widehat{L}_n^\gamma(f)$ and empirical covering numbers in the case when the empirical error equals zero. One novelty in our approach is that we appeal to general concentration-of-measure inequalities derived by Boucheron, Lugosi, and Massart (2000) in dealing with the empirical fat-shattering dimension. In a recent work Bartlett and Mendelson (2002) develop data-dependent bounds of the same kind as those offered in this paper. The bounds in Bartlett and Mendelson (2002) are based on Rademacher and Gaussian complexities and are used in deriving easy-to-compute bounds in some important special cases such as support vector machines and neural networks.

The results of Bartlett and Mendelson (2002) and of the present paper are not directly comparable. Our Theorem 3 is close in spirit to the basic results of Bartlett and Mendelson.

The rest of the paper is organized as follows. In Sections 2 and 3 we present the main data-dependent upper bounds for the probability of misclassification. In Section 4 we also develop an alternative data-based bound which provides a more easily computable data-dependent bound on the generalization error of classes of convex combinations of classifiers. In Section 5 we provide nontrivial examples of function classes for which the data-dependent quantities appearing in the main inequalities of Section 2 may be either computed exactly or bounded efficiently. Section 6 contains the proofs of the theorems.

2. Bounding by the random fat-shattering dimension

For $\gamma > 0$, a sequence $x_1^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$ is said to be γ-shattered by $\mathcal{F}$ if there is an $(r_1, \ldots, r_n) \in \mathbb{R}^n$ such that for each $(b_1, \ldots, b_n) \in \{0,1\}^n$ there is an $f \in \mathcal{F}$ satisfying for all $i = 1, \ldots, n$,

$$f(x_i) \geq r_i + \gamma \text{ if } b_i = 1, \text{ and } f(x_i) \leq r_i - \gamma \text{ if } b_i = 0$$

or, equivalently,

$$(2b_i - 1)(f(x_i) - r_i) \geq \gamma. \tag{1}$$

The (empirical) fat-shattering dimension (γ-dimension) of $\mathcal{F}$ in a sequence $x_1^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$ is defined for any $\gamma > 0$ by

$$\mathrm{fat}_{\mathcal{F}, x_1^n}(\gamma) = \max\{m : \mathcal{F}\ \gamma\text{-shatters a subsequence of length } m \text{ of } x_1^n\}.$$

Note that for $X_1^n = (X_1, \ldots, X_n)$, $\mathrm{fat}_{\mathcal{F}, X_1^n}(\gamma)$ is a random quantity whose value depends on the data. The (worst-case) fat-shattering dimension

$$\overline{\mathrm{fat}}_{\mathcal{F},n}(\gamma) = \sup_{x_1^n \in \mathcal{X}^n} \mathrm{fat}_{\mathcal{F}, x_1^n}(\gamma)$$

was used by Kearns and Schapire (1994), Alon et al. (1997), Shawe-Taylor et al. (1998), and Bartlett (1998) to derive useful bounds. In particular, Anthony and Bartlett (1999) show that if $d = \overline{\mathrm{fat}}_{\mathcal{F},n}(\gamma/8)$, then for any $0 < \delta < 1/2$, with probability at least $1 - \delta$, all $f \in \mathcal{F}$ satisfies

$$L(f) < \widehat{L}_n^\gamma(f) + 2.829\sqrt{\frac{1}{n}\left(d \log_2\left(\frac{32en}{d}\right)\ln(128n)\right)} + 2.829\sqrt{\frac{\ln(4/\delta)}{n}}\,. \tag{2}$$

(Throughout this paper $\log_b$ denotes the logarithm to the base b and ln denotes the natural logarithm.)

Before stating the first two main theorems, we need to introduce the notion of covering and packing numbers. Let (S, ρ) be a metric space. For $\epsilon > 0$, the ϵ-covering number $N_\rho(\epsilon, S)$ of S is defined as the minimum number of open balls of radius ϵ in S whose union covers S. (If no such finite cover exists, we formally define $N_\rho(\epsilon, S) = \infty$.)

A set $W \subset S$ is said to be ϵ-separated if $\rho(x, y) \geq \epsilon$ for all distinct $x, y \in W$. The ϵ-packing number $M_\rho(\epsilon, S)$ is defined as the maximum cardinality of an ϵ separated subset of S.

For $x_1^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and a family $\mathcal{G}$ of functions mapping $\mathcal{X}$ into $\mathbb{R}$, let $\mathcal{G}_{x_1^n}$ denote the subset of $\mathbb{R}^n$ given by

$$\mathcal{G}_{x_1^n} = \{(g(x_1), \ldots, g(x_n)) : g \in \mathcal{G}\}.$$

Let ρ_∞ denote the l_∞ metric on $\mathbb{R}^n$, given for any $u_1^n, v_1^n \in \mathbb{R}^n$ by $\rho_\infty(u_1^n, v_1^n) = \max_{1 \le i \le n} |u_i - v_i|$ and, for $\epsilon > 0$ and $x_1^n \in \mathcal{X}^n$, define

$$\mathcal{N}_\infty(\epsilon, \mathcal{G}, x_1^n) = N_{\rho_\infty}(\epsilon, \mathcal{G}_{x_1^n})$$

and

$$\mathcal{M}_\infty(\epsilon, \mathcal{G}, x_1^n) = M_{\rho_\infty}(\epsilon, \mathcal{G}_{x_1^n}).$$

The next result is an improvement over (2) in that we are able to replace the worst-case fat-shattering dimension $\overline{\mathrm{fat}}_{\mathcal{F},n}(\gamma/8)$ by its empirical counterpart $\mathrm{fat}_{\mathcal{F},X_1^n}(\gamma/8)$. Since for certain "lucky" distributions of the data the improvement is significant, such an empirical bound can play a crucial role in model selection.

Theorem 1 *Let $\mathcal{F}$ be a class of real measurable functions on $\mathcal{X}$, let $\gamma > 0$, and set $d(X_1^n) = \mathrm{fat}_{\mathcal{F},X_1^n}(\gamma/8)$. Then for any $0 < \delta < 1$, the probability that all $f \in \mathcal{F}$ satisfy*

$$L(f) \le \widehat{L}_n^\gamma(f) + \sqrt{\frac{1}{n}\left(9d(X_1^n) + 12.5 \ln\frac{8}{\delta}\right) \ln\left(\frac{32en}{d(X_1^n)}\right) \ln(128n)}$$

is greater than $1 - \delta$.

The following result improves Theorem 1 if $\widehat{L}_n^\gamma(f)$ is very small.

Theorem 2 *Consider the notation of Theorem 1. Then for any $0 < \delta < 1$, the probability that all $f \in \mathcal{F}$ satisfy*

$$L(f) \le \inf_{\alpha > 0}\left[(1+\alpha)\,\widehat{L}_n^\gamma(f) + \frac{1+\alpha}{n\alpha}\left(18d(X_1^n) + 25 \ln\frac{8}{\delta}\right) \ln\left(\frac{32en}{d(X_1^n)}\right) \ln(128n)\right]$$

is greater than $1 - \delta$.

The proofs of Theorems 1 and 2 are found in Section 6.

The result of Shawe-Taylor and Williamson (1999) assumes $\widehat{L}_n^\gamma(f) = 0$ and in that case states an inequality similar to the second inequality of Theorem 2.

Remark. As a reviewer pointed out to us, the factor $(1+\alpha)/\alpha$ in the second term of the upper bound in the theorem may be replaced by $(1 + 4\alpha + \alpha^2)/(4\alpha)$. This improves the bound by a constant factor whenever $\alpha < \sqrt{3}$.

3. An alternative data-based bound

In this section we propose another new data-dependent upper bound for the probability of error. The estimate is close, in spirit, to the recently introduced estimates of Koltchinskii and Panchenko (2002) based on Rademacher complexities, and the maximum discrepancy estimate of Bartlett, Boucheron, and Lugosi (2001).

Assume that n is even, and, for each $f \in \mathcal{F}$, consider the empirical error

$$\widehat{L}^{(2)}_{n/2}(f) = \frac{2}{n} \sum_{i=n/2+1}^{n} I_{\{\mathrm{sgn}(f(X_i)-1/2) \neq Y_i\}}$$

measured on the second half of the data. This may be compared with the sample error of f, with respect to margin γ, measured on the first half of the data

$$\widehat{L}^{\gamma}_{n/2}(f) = \frac{2}{n} \sum_{i=1}^{n/2} I_{\{\mathrm{margin}(f(X_i),Y_i) < \gamma\}} \ .$$

We have the following data-based estimate for the probability of error of any classifier in $\mathcal{F}$:

Theorem 3 *Let $\mathcal{F}$ be a class of real measurable functions on $\mathcal{X}$, let $\gamma > 0$. Then for any $0 < \delta < 1/2$, the probability that all $f \in \mathcal{F}$ satisfy*

$$L(f) < \widehat{L}^{\gamma}_{n}(f) + \sup_{f' \in \mathcal{F}} \left(\widehat{L}^{(2)}_{n/2}(f') - \widehat{L}^{\gamma}_{n/2}(f') \right) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}$$

is at least $1 - \delta$.

The proof is postponed to Section 6.

Remark. Theorem 3 is, modulo a small constant factor, always at least as good as Theorem 1. This may be seen by observing that by concentration of $\sup_{f \in \mathcal{F}} \left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f) \right)$ (which can be easily quantified using the bounded difference inequality (McDiarmid, 1989)),

$$\sup_{f \in \mathcal{F}} \left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f) \right) \approx \mathbf{E} \sup_{f \in \mathcal{F}} \left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f) \right)$$

with very large probability. By inspecting the proof of Theorem 1 it is easy to see that this expectation, with very large probability, does not exceed a quantity of the form

$$c\sqrt{\frac{1}{n} \left(d(X_1^n) + \ln \frac{1}{\delta} \right) \ln \left(\frac{32en}{d(X_1^n)} \right) \ln(n)}$$

for an appropriate constant c. The details are omitted.

4. Convex hulls

In this section we consider an important class of special cases. Let $\mathcal{H}$ be a class of "base" classifiers, that is, a class of functions $h : \mathcal{H} \to \{0,1\}$. Then we may define the class $\mathcal{F}$ of all (finite) convex combinations of elements of $\mathcal{H}$ by

$$\mathcal{F} = \left\{ f(x) = \sum_{i=1}^{N} w_i h_i(x) : N \geq 1, w_1, \ldots, w_N \geq 0, \sum_{i=1}^{N} w_i = 1, h_1, \ldots, h_N \in \mathcal{H} \right\}. \qquad (3)$$

Voting methods such as bagging and boosting choose a classifier from a class of classifiers of the above form. A practical disadvantage of the upper bounds appearing in Theorems 1 and 3 is that their computation may be prohibitively complex. For example, the bound of Theorem 3 involves optimization over the whole class $\mathcal{F}$. In the argument below we show, using ideas of Koltchinskii and Panchenko (2002), that at the price of weakening the bound of Theorem 3 we may obtain a data-dependent bound whose computation is significantly less complex than that of the bound of Theorem 3. Observe that to calculate the upper bound of the theorem below, it suffices to optimize over the "small" class of base classifiers $\mathcal{H}$.

Theorem 4 *Let $\mathcal{F}$ be a class of the form (3). Then for any $0 < \delta < 1/2$ and $\gamma > 0$, the probability that all $f \in \mathcal{F}$ satisfy*

$$L(f) < \widehat{L}_n^{\gamma}(f) + \frac{1}{\gamma} \sup_{h \in \mathcal{H}} \left(\widehat{L}_{n/2}^{(1)}(h) - \widehat{L}_{n/2}^{(2)}(h) \right) + \left(5 + \frac{2}{\gamma}\right) \sqrt{\frac{\ln(4/\delta)}{2n}}$$

is at least $1 - \delta$, where $\widehat{L}_{n/2}^{(1)}(h) = \frac{2}{n} \sum_{i=1}^{n/2} I_{\{\mathrm{sgn}(h(X_i) - 1/2) \neq Y_i\}}$.

The proof, based on arguments of Koltchinskii and Panchenko (2002) and concentration inequalities, is given in Section 6.

Remark. To interpret this new bound note that, for all $\delta > 0$, by the bounded difference inequality (McDiarmid, 1989), with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}} \left(\widehat{L}_{n/2}^{(1)}(h) - \widehat{L}_{n/2}^{(2)}(h) \right) \leq \mathbf{E} \sup_{h \in \mathcal{H}} \left(\widehat{L}_{n/2}^{(1)}(h) - \widehat{L}_{n/2}^{(2)}(h) \right) + \sqrt{\frac{2 \ln(1/\delta)}{n}}.$$

The expectation on the right-hand side may be further bounded by the Vapnik-Chervonenkis inequality (see Devroye and Lugosi, 2000, for this version):

$$\mathbf{E} \sup_{h \in \mathcal{H}} \left(\widehat{L}_{n/2}^{(1)}(h) - \widehat{L}_{n/2}^{(2)}(h) \right) \leq \sqrt{\frac{8 \mathbf{E} \log_2 S_{\mathcal{H}}(X_1^n)}{n}}$$

where $S_{\mathcal{H}}(X_1^n)$ is the random shatter coefficient, that is, the number of different ways the data points $X_1, \ldots, X_n$ can be classified by elements of the base class $\mathcal{H}$. We may convert this bound into another data-dependent bound by recalling that, by Boucheron et al. (2000, Theorem 4.2), $\log_2 S_{\mathcal{H}}(X_1^n)$ is strongly concentrated around its mean. Putting the pieces together, we obtain that, with probability at least $1 - \delta$, all $f \in \mathcal{F}$ satisfy

$$L(f) < \widehat{L}_n^{\gamma}(f) + \frac{4}{\gamma} \sqrt{\frac{\log_2 S_{\mathcal{H}}(X_1^n)}{n}} + \left(5 + \frac{12}{\gamma}\right) \sqrt{\frac{\ln(8/\delta)}{2n}}\ .$$

Remark. The bound of Theorem 4 may be significantly weaker than that of Theorem 3. As an example, consider the case when $\mathcal{X} = [0,1]$, and let $\mathcal{H}$ be the class of all indicator functions of intervals in $\mathbb{R}$. In this case, Anthony and Bartlett (1999, Theorem 12.11) shows that $\overline{\text{fat}}_{\mathcal{F},n}(\gamma) \leq 2/\gamma+1$, and therefore Theorem 1 (and even (2)) yields a bound of the order $O\left(\sqrt{\ln^2 n/(\gamma n)}\right)$. Thus, the dependence of the bound of Theorem 4 on γ is significantly worse than those of Theorems 1 and 3. This is the price we pay for computational feasibility. It is an interesting problem to determine the optimal dependence of the bounds on the margin parameter γ.

5. Examples

In this section we present three examples of function classes for which the empirical fat-shattering dimension may be either computed exactly or bounded efficiently.

5.1 Example 1: convex hulls of one-dimensional piecewise-linear sigmoids

Consider the problem of measuring the empirical fat-shattering dimension of a simple function class, the class of convex combinations of one-dimensional "piecewise-linear sigmoids" with bounded slope. Our results here show that, at least in one-dimension, it is possible to measure the empirical fat-shattering dimension in polynomial time, and that the empirical fat-shattering dimension measured on a given data set can be considerably lower than the worst-case fat-shattering dimension.

Consider the family $\mathcal{G}_\alpha$ of one-dimensional piecewise-linear sigmoids with bounded slope. Formally, for $x_a, x_b, y_a, y_b \in \mathbb{R}$ such that $x_a < x_b$, let

$$g^{(x_a,x_b,y_a,y_b)}(x) \quad = \quad \begin{cases} y_a & \text{if } x \leq x_a \\ y_b & \text{if } x \geq x_b \\ y_a + \frac{y_b - y_a}{x_b - x_a}(x - x_a) & \text{otherwise} \end{cases}$$

and let $\mathcal{G}_\alpha = \{g^{(x_a,x_b,y_a,y_b)} : \left|\frac{y_b-y_a}{x_b-x_a}\right| \leq 2\alpha\}$. Let $\mathcal{F}_\alpha$ be the set of functions constructed by (3) using $\mathcal{G}_\alpha$ as the set of base classifiers. The next lemma will serve as a basis for a constructive algorithm that can measure $\text{fat}_{\mathcal{F}_\alpha, x_1^n}(\gamma)$ on any data set $x_1^n = \{x_1, \ldots, x_n\} \subset \mathbb{R}$.

Lemma 5 *An ordered set $x_1^n = \{x_1, \ldots, x_n\} \subset \mathbb{R}, x_i < x_{i+1}, i = 1, \ldots, n-1$, is γ-shattered by $\mathcal{F}_\alpha$ if and only if*

$$\sum_{i=2}^{n} \frac{1}{d_i} \leq \frac{\alpha}{\gamma} \tag{4}$$

where $d_i = x_i - x_{i-1}$.

The proof of Lemma 5 is found in Section 6.

Lemma 5 shows that to find the empirical fat-shattering dimension of a data set x_1^n, we have to find the largest subset of x_1^n for which (4) holds. Suppose that the points of x_1^n are indexed in increasing order, and let $d_{ij} = x_i - x_j$. First consider the problem of finding a subsequence of x_1^n of length k that minimizes the cost $\sum_{i=1}^{k-1} \frac{1}{d_{j_{i+1},j_i}}$ over all subsequences of length k. Let $S(k;p,r) = (x_p = x_{j_1}, \ldots, x_{j_{k+1}} = x_r)$ denote the optimal subsequence of length $k+1$ between x_p and x_r, and let $C(k;p,r) = \sum_{i=1}^{k} \frac{1}{d_{j_i,j_{i+1}}}$ be the cost of $S(k;p,r)$. Observe that any subsequence $(x_{j_i}, \ldots, x_{j_{i+\ell-1}})$ of $S(k;p,r)$ of length ℓ is optimal over all subsequences of length ℓ between x_{j_i} and $x_{j_{i+\ell-1}}$, so $C(k;p,r)$ can be defined recursively as

$$C(k;p,r) = \begin{cases} \frac{1}{d_{p,r}} & \text{if } k = 1 \\ \min\limits_{q:p+k-1 \le q \le r-1} \big(C(k-1;p,q) + C(1;q,r)\big) & \text{if } k > 1. \end{cases}$$

Observe also that if $C(k-1;1,r)$ is known for all the $O(n)$ different indices r, then $C(k;1,r)$ can be calculated in $O(n^2)$ time for all r. Thus, by using a dynamic programming approach, we can find the sequence $C(1;1,n), C(2;1,n), \ldots, C(k;1,n)$ in $O(n^2 k)$ time. To compute $\mathrm{fat}_{\mathcal{F}_\alpha, x_1^n}(\gamma)$, notice that $\mathrm{fat}_{\mathcal{F}_\alpha, x_1^n}(\gamma) = k$ if and only if $C(k-1;1,n) \le \frac{\alpha}{\gamma}$ and either $C(k;1,n) > \frac{\alpha}{\gamma}$ or $k = n$. The algorithm is given formally in Figure 1.

```
FatLinearSigmoid(X, α, γ)

n ← X.length
for p ← 1 to n − 1 do
    for r ← p + 1 to n do
        C[1, p, r] ← 1 / (X[r] − X[p])
k ← 1
while C[k, 1, n] ≤ α/γ do
    k ← k + 1
    if k = n then
        return k
    for r ← k + 1 to n do
        C[k, 1, r] ← ∞
        for q ← k to r − 1 do
            c ← C[k − 1, 1, q] + C[1, q, r]
            if c < C[k, 1, r] then
                C[k, 1, r] ← c
return k
```

Figure 1: FATLINEARSIGMOID(X, α, γ) computes $\mathrm{fat}_{\mathcal{F}_\alpha, x_1^n}(\gamma)$ in $O(n^2 \, \mathrm{fat}_{\mathcal{F}_\alpha, x_1^n})$ time. The input array X contains the data points in increasing order.

It is clear from Lemma 5 that the worst-case fat-shattering dimension $\overline{\mathrm{fat}}_{\mathcal{F}_\alpha, n}(\gamma) = n$ for all $\gamma > 0$ if the data points may take any value in $\mathbb{R}$. Thus, the data-dependent dimension $\mathrm{fat}_{\mathcal{F}_\alpha, x_1^n}(\gamma)$ presents a qualitative improvement. If the data points $x_1, \ldots, x_n$ are restricted to fall in the an interval of length A then it follows from Lemma 5 and the inequality between arithmetic and harmonic means that $\overline{\mathrm{fat}}_{\mathcal{F}_\alpha, n}(\gamma) = \left\lfloor \sqrt{A\alpha/\gamma} \right\rfloor + 1$. This upper bound is achieved by equispaced data points. Even in this case, the empirical fat-shattering dimension may be significantly smaller than its worst-case upper bound, and the difference is larger if the data points are very unevenly distributed. To experimentally quantify this intuition, we compared the fat-shattering dimension of data sets drawn from different distributions over $[0, 1]$. Figure 2(a) shows that even in the case of uniform distribution, for high α/γ ratio we gain approximately 20% over the data-independent fat-shattering dimension. As the points become more and more unevenly distributed (Gaussian distributions with decreasing standard deviations), the difference between the data-independent and data-dependent fat-shattering dimensions increases.

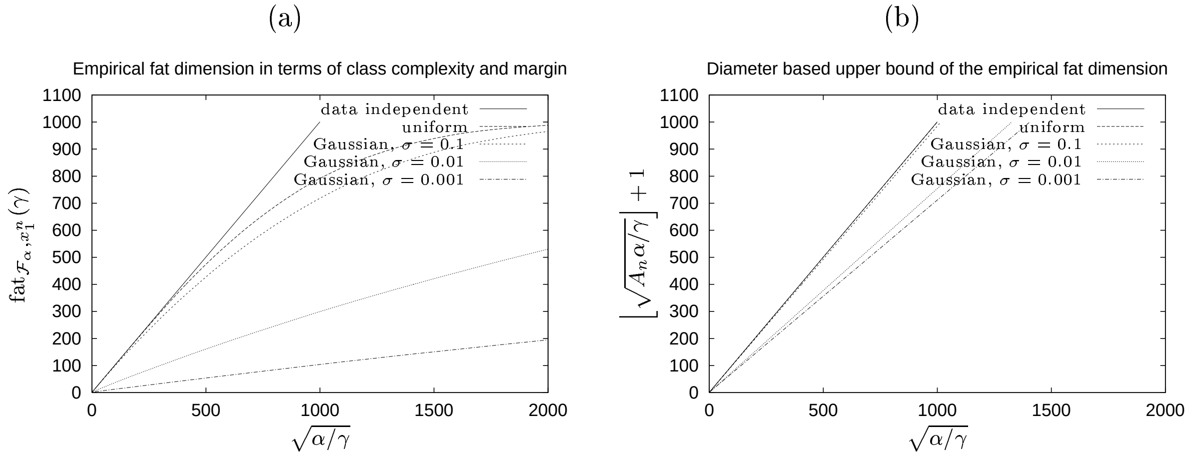

Figure 2: The first figure shows the empirical fat-shattering dimensions of different data sets as a function of the class complexity α and the margin γ. The second figure indicates the upper bound (5) based on the empirical diameter A_n of the data. The solid lines in both figures show the data-independent fat-shattering dimension $\overline{\mathrm{fat}}_{\mathcal{F}_\alpha, n}(\gamma) = \left\lfloor\sqrt{A\alpha/\gamma}\right\rfloor + 1$ achieved by equispaced data points. We generated data sets of 1000 points drawn from the uniform distribution in $[0, 1]$, and from the mixture of two identical Gaussians with means $1/4$ and $3/4$, and standard deviations indicated by the figure. The Gaussian mixtures were truncated to $[0, 1]$ to keep their data-independent fat-shattering dimension finite.

The empirical diameter $A_n = \max_i x_i - \min_i x_i$ can also be used to bound the data-dependent fat-shattering dimension from above since

$$\mathrm{fat}_{\mathcal{F}_\alpha, x_1^n}(\gamma) \leq \left\lfloor\sqrt{A_n\alpha/\gamma}\right\rfloor + 1. \tag{5}$$

The computation of (5) is, of course, trivial. Figure 2(b) shows that if the empirical diameter A_n is significantly smaller then the a-priori diameter A, the bound (5) can provide an improvement over the data-independent fat-shattering dimension. Such simple upper bounds for the empirical fat-shattering dimension may be useful in practice and may be easy to obtain in more general situations as well. However, if the data is unevenly distributed in the empirical support $[\min_i x_i, \max_i x_i]$, $\mathrm{fat}_{\mathcal{F}_\alpha, x_1^n}(\gamma)$ can be much smaller than the empirical diameter-based bound (5).

5.2 Example 2: one-dimensional piecewise-linear sigmoids with L_p constraint

In the practice of neural networks, the L_2 regularization constraint is more often considered than boosting's L_1 constraint mainly because it is easier to optimize an objective function with an L_2 constraint. Below we consider the general case of L_p regularization constraint. Interestingly, the empirical fat-shattering dimensions of such classes depend not only on the weight constraint but also on the number of neurons as we show below in a one-dimensional example.

Consider the class of linear combinations of N one-dimensional piecewise-linear sigmoids with an L_p constraint,

$$\mathcal{F}_{\alpha,N,p} = \left\{ f(x) = \sum_{i=1}^{N} w_i g_i(x) : w_1, \ldots, w_N \geq 0, \sum_{i=1}^{N} w_i^p \leq 1, g_1, \ldots, g_N \in \mathcal{G}_\alpha \right\}$$

where $p \geq 1$, and G_α is the same as in Example 1. First, observe that Jensen's inequality implies

$$\left(\frac{1}{N} \sum_{i=1}^{N} w_j \right)^p \leq \frac{1}{N} \sum_{i=1}^{N} w_j^p \leq \frac{1}{N},$$

so using the second half of the proof of Lemma 5 we can show the following.

Lemma 6 *If an ordered set $x_1^n = \{x_1, \ldots, x_n\} \subset \mathbb{R}, x_i < x_{i+1}, i = 1, \ldots, n-1$, is γ-shattered by $\mathcal{F}_{\alpha,N,p}$, then*

$$\sum_{i=2}^{n} \frac{1}{d_i} \leq \frac{\alpha}{\gamma} N^{\frac{p-1}{p}} \tag{6}$$

where $d_i = x_i - x_{i-1}$.

Unfortunately, we cannot prove the reverse statement, i.e., that (6) implies that $\mathcal{F}_{\alpha,N,p}$ γ-shatters x_1^n. However, the size of the largest subset of x_1^n for which (6) holds is an upper bound of $\text{fat}_{\mathcal{F}_{\alpha,N,p},x_1^n}(\gamma)$, so it can be used to upper bound the error probability in Theorem 1. To find the largest subset, we can use Algorithm FATLINEARSIGMOID with line 6 replaced by

6	**while** $C[k,1,n] \leq \frac{\alpha}{\gamma} N^{\frac{p-1}{p}}$ **do**

5.3 Example 3: multivariate Lipschitz functions

In this section we consider a simple multivariate function class, the class of Lipschitz functions

$$\text{Lip}_{2\alpha} = \{ f : \mathbb{R}^d \to \mathbb{R}, \forall x, y \in \mathbb{R}^d : \|f(x) - f(y)\| \leq 2\alpha \|x - y\| \}.$$

Although this function class is seldom used in practice, it has the same "flavor" as some function classes used in practical algorithms (such as the support vector machines or neural networks with weight decay) that control the capacity of the classifiers by implicitly constraining the slope of the underlying discriminant functions. Of course $\text{Lip}_{2\alpha}$ is easier to deal with since function classes used by practical algorithms tend to have data-dependent, non-uniform constraints on their slope.

We first show that the computation of the exact empirical fat-shattering dimension of this class is NP-hard. Then we describe a greedy approximation algorithm that computes lower and upper bounds of the empirical fat-shattering dimension in polynomial time. We demonstrate on real data sets that the upper bound can be by several orders of magnitude better then the data-independent fat-shattering dimension, especially if the data dimension is large.

To show that the computation of the exact empirical fat-shattering dimension of class $\mathrm{Lip}_{2\alpha}$ is NP-hard, we rely on the following simple fact.

Lemma 7 *A set $x_1^n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, is γ-shattered by $\mathrm{Lip}_{2\alpha}$ if and only if no two points in x_1^n are closer to each other than $\frac{\gamma}{\alpha}$.*

Proof. By the definition of $\mathrm{Lip}_{2\alpha}$, if two points are closer than $\frac{\gamma}{\alpha}$, then no $f \in \mathrm{Lip}_{2\alpha}$ can separate the two points with different labels by a margin of 2γ. On the other hand, if no two points in x_1^n are closer to each other than $\frac{\gamma}{\alpha}$, then for any labeling $\{y_1, \ldots, y_n\} \in \{0,1\}^n$ we can construct a γ-separating function in $\mathrm{Lip}_{2\alpha}$ in the following way. For each $x_i \in x_1^n$ let

$$f_i(x) = \begin{cases} (2y_i - 1)(\gamma - \|x - x_i\| 2\alpha) & \text{if } \|x - x_i\| \leq \frac{\gamma}{2\alpha}, \\ 0 & \text{otherwise,} \end{cases}$$

and let $f(x) = \sum_{i=1}^n f_i(x)$. Each f_i is in $\mathrm{Lip}_{2\alpha}$, and at every point $x \in \mathbb{R}^d$ at most one function f_i can take a nonzero value so $f \in \mathrm{Lip}_{2\alpha}$. At the same time, since $f(x_i) = f_i(x_i) = (2y_i - 1)\gamma$, f γ-separates x_1^n. $\square$

Thus, to find the empirical fat-shattering dimension $\mathrm{fat}_{\mathrm{Lip}_{2\alpha}, x_1^n}(\gamma)$ we have to find the size of the largest subset of x_1^n that satisfies the condition of Lemma 7. To this end, we define a graph $G_{\alpha,\gamma}(V, E)$ where $V = x_1^n$ and two points are connected with an edge if and only if they are closer to each other than $\frac{\gamma}{\alpha}$. Finding the size of the largest subset of x_1^n that satisfies the condition of Lemma 7 is equivalent to finding the size of a maximum independent vertex set $\textsc{MaxInd}(G)$ of $G_{\alpha,\gamma}$, which is an NP-hard problem. There are results that show that for a general graph, even the approximation of $\textsc{MaxInd}(G)$ within a factor of $n^{1-\epsilon}$, for any $\epsilon > 0$, is NP-hard (Hastad, 1996). On the positive side, it was shown that for such geometric graphs as $G_{\alpha,\gamma}$, $\textsc{MaxInd}(G)$ can be approximated arbitrarily well by polynomial time algorithms (Erlebach et al., 2001). However, approximating algorithms of this kind scale exponentially with the data dimension both in terms of the quality of the approximation and the running time[1] so they are of little practical use for $d > 2$. Hence, instead of using these algorithms, we apply a greedy approximation method that provides lower and upper bounds of $\mathrm{fat}_{\mathrm{Lip}_{2\alpha}, x_1^n}(\gamma)$ in polynomial time in both n and d.

The approach is based on the basic relation between packing and covering numbers. Let $\mathcal{N}(r, x_1^n)$ be the smallest subset of x_1^n such that for every $x \in x_1^n$ there exists a point c (called center) in $\mathcal{N}(r, x_1^n)$ such that $\|x - c\| \leq r$, and let $\mathcal{M}(r, x_1^n)$ be the largest subset of

1. Typically, the computation of an independent vertex set of G of size at least $\left(1 - \frac{1}{k}\right)^d \textsc{MaxInd}(G)$ requires $O(n^{k^d})$ time.

x_1^n such that for every $c_1, c_2 \in \mathcal{M}(r, x_1^n)$, $\|c_1 - c_2\| > r$. For notational simplicity we will denote $\mathcal{N}(r, x_1^n)$ and $\mathcal{M}(r, x_1^n)$ by $\mathcal{N}_r$ and $\mathcal{M}_r$, respectively. Let $N_r = |\mathcal{N}_r|$ and $M_r = |\mathcal{M}_r|$. By the definition of M_r and $\mathrm{fat}_{\mathrm{Lip}_{2\alpha}, x_1^n}(\gamma)$ it is clear that $\mathrm{fat}_{\mathrm{Lip}_{2\alpha}, x_1^n}(\gamma) = M_{\gamma/\alpha}$, and it is well known that $M_r \le N_{r/2}$. To find the exact values of M_r and $N_{r/2}$ is a hard problem. However, the size of any r-packing is a lower bound to M_r, and the size of any $\frac{r}{2}$-covering is an upper bound. To compute these lower and upper bounds, we designed two algorithms that, for each r_i in a predefined sequence $r_1, r_2, \ldots, r_N$, construct an r_i-packing and an $\frac{r_i}{2}$-covering, respectively, of x_1^n. We omit the details of these algorithms but show on Figure 3 the results on four datasets[2] from the UCI data repository (Blake et al., 1998).

It is clear from Lemma 7 that the worst-case fat-shattering dimension $\mathrm{fat}_{\mathrm{Lip}_{2\alpha}, n}(\gamma) = n$ for all $\gamma > 0$ if the data points may take any value in $\mathbb{R}^d$. If the support S of the data distribution is finite, then $\mathrm{fat}_{\mathrm{Lip}_{2\alpha}, n}(\gamma)$ is equal to the $\frac{\gamma}{\alpha}$-packing number of S which can still be by several orders of magnitude larger than the data-independent fat-shattering dimension, especially if the data dimension is large.

6. Proofs

The main ideas behind the the proofs of Theorems 1 and 2 are rather similar. Both proofs use, in a crucial way, the fact that the empirical fat shattering dimension is sharply concentrated around its expected value. However, to obtain the best bounds, the usual symmetrization steps need to be revisited and appropriate modifications have to be made. We give the somewhat more involved proof of Theorem 2 in detail, and only indicate the main steps of the proof of Theorem 1. In both proofs we let (X_i, Y_i), $i = n+1, \ldots, 2n$, be i.i.d. copies of (X, Y), independent of D_n, and define, for each $f \in \mathcal{F}$,

$$\widehat{L'}_n(f) = \frac{1}{n} \sum_{i=n+1}^{2n} I_{\{\mathrm{sgn}(f(X_i) - 1/2) \ne Y_i\}} \quad \text{and} \quad \widehat{L'}_n^{\gamma}(f) = \frac{1}{n} \sum_{i=n+1}^{2n} I_{\{\mathrm{margin}(f(X_i), Y_i) < \gamma\}}.$$

Proof of Theorem 2

Step 1 *For any positive measurable function $\epsilon(X_1^n)$ of X_1^n,*

$$\mathbf{P}\left\{\exists f \in \mathcal{F} : L(f) > \inf_{\alpha > 0}\left[(1+\alpha)\,\widehat{L}_n^{\gamma}(f) + \epsilon^2(X_1^n)\frac{1+\alpha}{\alpha}\right]\right\} \le \mathbf{P}\left\{\sup_{f \in \mathcal{F}} \frac{L(f) - \widehat{L}_n^{\gamma}(f)}{\sqrt{L(f)}} > \epsilon(X_1^n)\right\}$$

Proof. Assume that the event $\sup_{f \in \mathcal{F}}(L(f) - \widehat{L}_n^{\gamma}(f))/\sqrt{L(f)} > \epsilon(X_1^n)$ does not occur. Then for all $f \in \mathcal{F}$, we have $L(f) - \widehat{L}_n^{\gamma}(f) \le \epsilon(X_1^n)\sqrt{L(f)}$. There are two cases. Either

2. In a preprocessing step, categorical attributes were binary coded in a 1-out-of-n fashion. Data points with missing attributes were removed. Each attribute was normalized to have zero mean and $1/\sqrt{d}$ standard deviation. The four data sets were the Wisconsin breast cancer ($n = 683$, $d = 9$), the ionosphere ($n = 351$, $d = 34$), the Japanese credit screening ($n = 653$, $d = 42$), and the tic-tac-toe endgame ($n = 958$, $d = 27$) database.

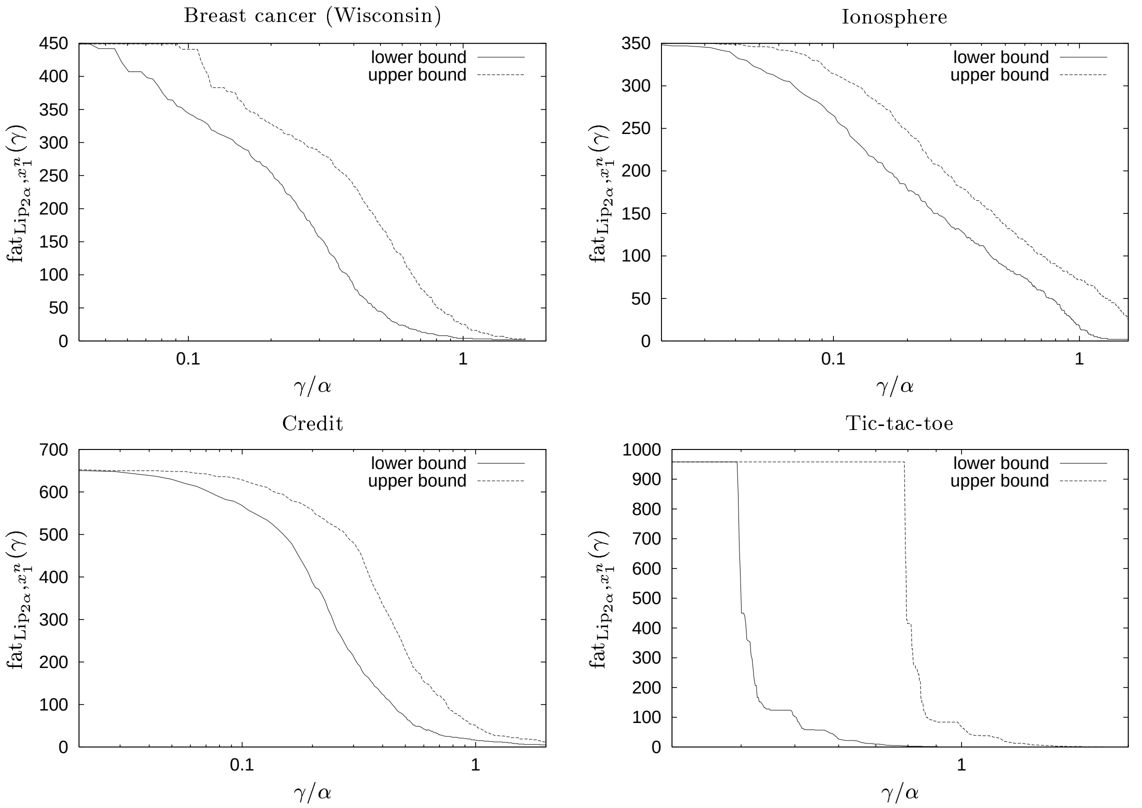

Figure 3: Upper and lower bounds of $\mathrm{fat}_{\mathrm{Lip}_{2\alpha}, x_1^n}(\gamma)$ for four datasets from the UCI data repository.

$f \in \mathcal{F}$ is such that $L(f) < (1 + 1/\alpha)^2 \epsilon(X_1^n)^2$ or $L(f) \geq (1 + 1/\alpha)^2 \epsilon(X_1^n)^2$. In the first case,

$$L(f) \leq \widehat{L}_n^\gamma(f) + (1 + 1/\alpha)\epsilon(X_1^n)^2 .$$

In the second case $L(f) \leq \widehat{L}_n^\gamma(f) + L(f)/(1 + 1/\alpha)$, which, after rearranging, implies

$$L(f) \leq \widehat{L}_n^\gamma(f)(1 + \alpha).$$

Thus, for every $f \in \mathcal{F}$,

$$L(f) \leq \inf_{\alpha > 0} \left[\widehat{L}_n^\gamma(f)(1 + \alpha) + (1 + 1/\alpha)\epsilon(X_1^n)^2 \right]$$

which implies the statement.

Step 2 *For any $n \geq 1$ and measurable function $\epsilon(X_1^n)$ of X_1^n such that $n\epsilon^2(X_1^n) \geq 2$ with probability one,*

$$\mathbf{P}\left\{ \sup_{f \in \mathcal{F}} \frac{L(f) - \widehat{L}_n^\gamma(f)}{\sqrt{L(f)}} > \epsilon(X_1^n) \right\} \leq 4\mathbf{P}\left\{ \sup_{f \in \mathcal{F}} \frac{\widehat{L}'_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{(\widehat{L}'_n(f) + \widehat{L}_n^\gamma(f))/2}} > \epsilon(X_1^n) \right\} .$$

Proof. Define $\mathcal{F}'$, a random subset of $\mathcal{F}$, by

$$\mathcal{F}' = \mathcal{F}'(X_1^n) = \left\{ f \in \mathcal{F} : L(f) - \widehat{L}_n^\gamma(f) > \epsilon(X_1^n)\sqrt{L(f)} \right\}$$

and note that $I_{\{\mathcal{F}'=\emptyset\}}$ and X_{n+1}^{2n} are independent. Observe that if $f \in \mathcal{F}'$ (implying $L(f) > \epsilon^2(X_1^n) > 0$) and additionally $\widehat{L'}_n(f) \geq L(f)$ (implying also $\widehat{L'}_n(f) > \widehat{L}_n^\gamma(f)$), then

$$\frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{(\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f))/2}} \geq \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f)}} = \sqrt{\widehat{L'}_n(f)} - \frac{\widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f)}} \geq \sqrt{L(f)} - \frac{\widehat{L}_n^\gamma(f)}{\sqrt{L(f)}} > \epsilon(X_1^n).$$

On the other hand, conditioning on X_1^n, for $f \in \mathcal{F}'$ (using that $nL(f) > n\epsilon^2(X_1^n) \geq 2$) it is known that $\mathbf{P}\{\widehat{L'}_n(f) > L(f) | X_1^n\} \geq 1/4$ (Slud, 1977, see, e.g.,)). Thus

$$\begin{aligned}
\mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{(\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f))/2}} > \epsilon(X_1^n) \right\} &\geq \mathbf{P}\{\exists f \in \mathcal{F}' : \widehat{L'}_n(f) > L(f)\} \\
&= \mathbf{E}\left[\mathbf{P}\{\exists f \in \mathcal{F}' : \widehat{L'}_n(f) > L(f) | X_1^n\}\right] \\
&\geq \mathbf{E}\left[I_{\{\mathcal{F}'\neq\emptyset\}} \sup_{f\in\mathcal{F}'} \mathbf{P}\{\widehat{L'}_n(f) > L(f) | X_1^n\}\right] \\
&\geq \frac{1}{4}\mathbf{P}\{\mathcal{F}' \neq \emptyset\} \\
&= \frac{1}{4}\mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{L(f) - \widehat{L}_n^\gamma(f)}{\sqrt{L(f)}} > \epsilon(X_1^n) \right\}.
\end{aligned}$$

Step 3 *For any* $\epsilon > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\left(\frac{9}{n}d(X_1^n) + \frac{\epsilon^2}{2}\right) \ln\left(\frac{32en}{d(X_1^n)}\right) \ln(128n)} \right\} \\
&\quad \leq \mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\left(\frac{3}{n}d(X_1^{2n}) + \frac{\epsilon^2}{4}\right) \ln\left(\frac{32en}{d(X_1^{2n})}\right) \ln(128n)} \right\} + e^{-\frac{n\epsilon^2}{25}}
\end{aligned}$$

Proof. Define the event $A = \{d(X_1^{2n}) > 3d(X_1^n) + n\epsilon^2/12\}$. Then we can write

$$\begin{aligned}
&\mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\left(\frac{9}{n}d(X_1^n) + \frac{\epsilon^2}{2}\right) \ln\left(\frac{32en}{d(X_1^n)}\right) \ln(128n)} \right\} \\
&\leq \mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\left(\frac{9}{n}d(X_1^n) + \frac{\epsilon^2}{2}\right) \ln\left(\frac{32en}{d(X_1^n)}\right) \ln(128n)}\ ,\ A^c \right\} + \mathbf{P}\{A\} \\
&\leq \mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\left(\frac{3}{n}d(X_1^{2n}) + \frac{\epsilon^2}{4}\right) \ln\left(\frac{32en}{d(X_1^{2n})}\right) \ln(128n)}\ ,\ A^c \right\} + \mathbf{P}\{A\}
\end{aligned}$$

$$\leq \mathbf{P}\left\{\sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\left(\frac{3}{n}d(X_1^{2n}) + \frac{\epsilon^2}{4}\right)\ln\left(\frac{32en}{d(X_1^{2n})}\right)\ln(128n)}\right\} + \mathbf{P}\{A\}.$$

It remains to show that $\mathbf{P}\{A\} \leq e^{-\frac{n\epsilon^2}{25}}$. If A occurs, then, letting $M = n\epsilon^2/12$,

$$3d(X_1^n) + M < d(X_1^{2n}) \leq d(X_1^n) + d(X_{n+1}^{2n}),$$

and hence $d(X_{n+1}^{2n}) > 2d(X_1^n) + M$. Moreover, by Chernoff's bounding method for any $\lambda > 0$,

$$\begin{aligned}
\mathbf{P}\{A\} \leq \mathbf{P}\{d(X_{n+1}^{2n}) > 2d(X_1^n) + M\} &= \mathbf{P}\{e^{\lambda(d(X_{n+1}^{2n}) - 2d(X_1^n) - M)} > 1\} \\
&\leq \mathbf{E}\left[e^{\lambda(d(X_{n+1}^{2n}) - 2d(X_1^n) - M)}\right] \\
&= \mathbf{E}\left[e^{\lambda d(X_{n+1}^{2n})}\right]\mathbf{E}\left[e^{-2\lambda d(X_1^n)}\right]e^{-\lambda M} \\
&= \mathbf{E}\left[e^{\lambda d(X_1^n)}\right]\mathbf{E}\left[e^{-2\lambda d(X_1^n)}\right]e^{-\lambda M}, \qquad (7)
\end{aligned}$$

since $X_1, \ldots, X_{2n}$ are i.i.d. The random fat-shattering dimension $d(X_1^n)$ is a configuration function in the sense of Boucheron et al. (2000, Section 3), and therefore it satisfies the concentration inequality given in equation (18) of Boucheron et al. (2000): for any $\lambda \in \mathbb{R}$,

$$\ln \mathbf{E}\left[e^{\lambda(d(X_1^n) - \mathbf{E}d(X_1^n))}\right] \leq \mathbf{E}\left[d(X_1^n)\right](e^\lambda - \lambda - 1).$$

This and (7) imply

$$\begin{aligned}
\mathbf{P}\{A\} &\leq \mathbf{E}\left[e^{\lambda d(X_1^n)}\right]\mathbf{E}\left[e^{-2\lambda d(X_1^n)}\right]e^{-\lambda M} \leq e^{(e^\lambda - 1)\mathbf{E}d(X_1^n)}e^{(e^{-2\lambda} - 1)\mathbf{E}d(X_1^n)}e^{-\lambda M} \\
&= e^{(e^\lambda + e^{-2\lambda} - 2)\mathbf{E}d(X_1^n) - \lambda M} = e^{-\ln(\frac{\sqrt{5}+1}{2})M} < e^{-\ln(\frac{\sqrt{5}+1}{2})\frac{n\epsilon^2}{12}} \\
&< e^{-\frac{n\epsilon^2}{25}},
\end{aligned}$$

where in the second equality we set $\lambda = \ln(\frac{\sqrt{5}+1}{2})$ so that $e^\lambda + e^{-2\lambda} - 2 = 0$.

Step 4 *For $\gamma > 0$, let $\pi_\gamma : \mathbb{R} \to [1/2 - \gamma, 1/2 + \gamma]$ be the "hard-limiter" function*

$$\pi_\gamma(t) = \begin{cases} 1/2 - \gamma & \text{if } t \leq 1/2 - \gamma \\ t & \text{if } 1/2 - \gamma < t < 1/2 + \gamma \\ 1/2 + \gamma & \text{if } t \geq 1/2 + \gamma \end{cases}$$

and set $\pi_\gamma(\mathcal{F}) = \{\pi_\gamma \circ f : f \in \mathcal{F}\}$. Then for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\left(\frac{3}{n}d(X_1^{2n}) + \frac{\epsilon^2}{4}\right)\ln\left(\frac{32en}{d(X_1^{2n})}\right)\ln(128n)}\right\} \\
&\qquad \leq \mathbf{P}\left\{\sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\frac{2}{n}\ln \mathcal{N}_\infty(\gamma/2, \pi_\gamma(\mathcal{F}), X_1^{2n}) + 4\epsilon^2}\right\}. (8)
\end{aligned}$$

Proof. The probability on the left hand side is upper bounded by

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{\widehat{L'}_n(f)-\widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f)+\widehat{L}_n^\gamma(f)}} > \sqrt{\frac{3}{n}d(X_1^{2n})\ln\left(\frac{32en}{d(X_1^{2n})}\right)\ln(128n)+4\epsilon^2}\right\},$$

since $\ln(\frac{32en}{d(X_1^{2n})})\ln(128n)\geq\ln(16e)\ln(128)>16$.

The following upper bound on the random covering number of $\pi_\gamma(\mathcal{F})$ in terms of the random fat-shattering dimension of $\mathcal{F}$ is given in Anthony and Bartlett (1999, Theorem 12.13):

$$\mathcal{N}_\infty(\gamma/2,\pi_\gamma(\mathcal{F}),X_1^n)\leq 2(64n)^{d(X_1^n)\log_2(16en/d(X_1^n))}. \tag{9}$$

It is easy to see that (9) implies

$$2\ln\mathcal{N}_\infty(\gamma/2,\pi_\gamma(\mathcal{F}),X_1^{2n})\leq 3d(X_1^{2n})\ln\left(\frac{32en}{d(X_1^{2n})}\right)\ln(128n)$$

and hence (8) holds.

Step 5 *Suppose $\mathcal{G}$ is a minimal $\gamma/2$-cover of $\pi_\gamma(\mathcal{F})$ and $\sigma_1,\ldots,\sigma_n$ are i.i.d. Rademacher random variables (i.e., $\mathbf{P}\{\sigma_1=1\}=\mathbf{P}\{\sigma_1=-1\}=1/2$) which are also independent of $(X_i,Y_i)_{i=1}^{2n}$. Then for any positive measurable function $\beta(X_1^{2n})$ of X_1^{2n} that depends only on the set $\{X_1,\ldots,X_{2n}\}$,*

$$\begin{aligned}&\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{\widehat{L'}_n(f)-\widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f)+\widehat{L}_n^\gamma(f)}}>\beta(X_1^{2n})\right\}\\ \leq\ &\mathbf{P}\left\{\max_{g\in\mathcal{G}}\frac{\sum_{i=1}^n\sigma_i(I_{n+i}^{\gamma/2}(g)-I_i^{\gamma/2}(g))}{\sqrt{w(g)}}>\sqrt{n}\beta(X_1^{2n})\right\},\end{aligned}$$

where $I_i^\gamma(g)=I_{\{\mathrm{margin}(g(X_i),Y_i)<\gamma\}}$ and $w(g)=\sum_{i=1}^n\left|I_{n+i}^{\gamma/2}(g)-I_i^{\gamma/2}(g)\right|$.

Proof. Recall that $\mathcal{G}$ is a set of functions with $|\mathcal{G}|=\mathcal{N}_\infty(\gamma/2,\pi_\gamma(\mathcal{F}),X_1^{2n})$ elements such that for any $f\in\mathcal{F}$ there exists a $g\in\mathcal{G}$ such that $\max_{i\leq 2n}|\pi_\gamma(f(X_i))-g(X_i)|<\gamma/2$. Observe that if f and g are such that $\max_{i\leq 2n}|\pi_\gamma(f(X_i))-g(X_i)|<\gamma/2$, then

$$I_{\{\mathrm{sgn}(f(X_i)-1/2)\neq Y_i\}}\leq I_{\{\mathrm{margin}(g(X_i),Y_i)<\gamma/2\}}\ ,\quad i=1,\ldots,2n$$

and

$$I_{\{\mathrm{margin}(f(X_i),Y_i)<\gamma\}}\geq I_{\{\mathrm{margin}(g(X_i),Y_i)<\gamma/2\}}\ ,\quad i=1,\ldots,2n\ ,$$

which imply $\widehat{L'}_n(f)\leq\widehat{L'}_n^{\gamma/2}(g)$ and $\widehat{L}_n^\gamma(f)\geq\widehat{L}_n^{\gamma/2}(g)$. Thus, noting that the function $F(x,y)=(x-y)/\sqrt{x+y}$ is increasing in x and decreasing in y for $x\geq 0$ and $y\geq 0$, we also have

$$\frac{\widehat{L'}_n(f)-\widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f)+\widehat{L}_n^\gamma(f)}}\leq\frac{\widehat{L'}_n^{\gamma/2}(g)-\widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n^{\gamma/2}(g)+\widehat{L}_n^\gamma(f)}}\leq\frac{\widehat{L'}_n^{\gamma/2}(g)-\widehat{L}_n^{\gamma/2}(g)}{\sqrt{\widehat{L'}_n^{\gamma/2}(g)+\widehat{L}_n^{\gamma/2}(g)}}.$$

Thus, with $I_i^{\gamma}(g)$ and $w(g)$ as above, we see that

$$
\begin{aligned}
&\mathbf{P}\left\{\sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f)-\widehat{L}_n^{\gamma}(f)}{\sqrt{\widehat{L'}_n(f)+\widehat{L}_n^{\gamma}(f)}} > \beta(X_1^{2n})\right\} \\
&\qquad \le \mathbf{P}\left\{\max_{g\in\mathcal{G}} \frac{\widehat{L'}_n^{\gamma/2}(g)-\widehat{L}_n^{\gamma/2}(g)}{\sqrt{\widehat{L'}_n^{\gamma/2}(g)+\widehat{L}_n^{\gamma/2}(g)}} > \beta(X_1^{2n})\right\} \\
&\qquad = \mathbf{P}\left\{\max_{g\in\mathcal{G}} \frac{\frac{1}{n}\sum_{i=1}^n (I_{n+i}^{\gamma/2}(g)-I_i^{\gamma/2}(g))}{\sqrt{\frac{1}{n}\sum_{i=1}^n (I_{n+i}^{\gamma/2}(g)+I_i^{\gamma/2}(g))}} > \beta(X_1^{2n})\right\} \\
&\qquad \le \mathbf{P}\left\{\max_{g\in\mathcal{G}} \frac{\sum_{i=1}^n (I_{n+i}^{\gamma/2}(g)-I_i^{\gamma/2}(g))}{\sqrt{w(g)}} > \sqrt{n}\beta(X_1^{2n})\right\}.
\end{aligned}
$$

To finish the proof, observe that the last probability does not change if for $i \le n$, (X_i, Y_i) is exchanged with (X_{n+i}, Y_{n+1}). In particular, if $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables which are also independent of $(X_i, Y_i)_{i=1}^{2n}$, then the last probability equals

$$
\mathbf{P}\left\{\max_{g\in\mathcal{G}} \frac{\sum_{i=1}^n \sigma_i(I_{n+i}^{\gamma/2}(g)-I_i^{\gamma/2}(g))}{\sqrt{w(g)}} > \sqrt{n}\beta(X_1^{2n})\right\}.
$$

Step 6 *Set*

$$
\beta(X_1^n) = \sqrt{\frac{2}{n}\ln \mathcal{N}_\infty(\gamma/2, \mathcal{H}, X_1^{2n}) + 4\epsilon^2},
$$

where $\epsilon > 0$. *Then*

$$
\mathbf{P}\left\{\max_{g\in\mathcal{G}} \frac{\sum_{i=1}^n \sigma_i(I_{n+i}^{\gamma/2}(g)-I_i^{\gamma/2}(g))}{\sqrt{w(g)}} > \sqrt{n}\beta(X_1^{2n})\right\} \le e^{-2n\epsilon^2}. \tag{10}
$$

Proof. We need the following well-known lemma:

Lemma 8 *Let* $\sigma > 0$, $N \ge 2$, *and let* $Z_1, \ldots, Z_N$ *be real-valued random variables such that for all* $s > 0$ *and* $1 \le i \le N$, $\mathbf{E}\left[e^{sZ_i}\right] \le e^{s^2\sigma^2/2}$. *Then*

$$
\mathbf{P}\left\{\max_{i\le N} Z_i > \epsilon\right\} \le N e^{-\epsilon^2/2\sigma^2}.
$$

Define the zero-mean random variables

$$
Z(g) = \frac{\sum_{i=1}^n \sigma_i(I_{n+i}^{\gamma/2}(g)-I_i^{\gamma/2}(g))}{\sqrt{w(g)}},
$$

and observe that, by Hoeffding's inequality (Hoeffding, 1963), for any $s > 0$ and $g \in \mathcal{G}$, $Z(g)$ satisfies

$$
\mathbf{E}\left[e^{sZ(g)}\middle|(X_i,Y_i)_{i=1}^{2n}\right] = \prod_{i=1}^n \mathbf{E}\left[e^{s\frac{\sigma_i(I_{n+i}^{\gamma/2}(g)-I_i^{\gamma/2}(g))}{\sqrt{w(g)}}}\middle|(X_i,Y_i)_{i=1}^{2n}\right]
$$

$$\leq \left(e^{s^2/2w(g)}\right)^{w(g)} \cdot 1^{n-w(g)} = e^{s^2/2}.$$

The proof is finished by applying Lemma 8 to the $|\mathcal{G}| = \mathcal{N}_\infty(\gamma/2, \pi_\gamma(\mathcal{F}), X_1^{2n})$ random variables $\{Z(g) : g \in \mathcal{G}\}$, if we first condition on $(X_i, Y_i)_{i=1}^{2n}$:

$$\begin{aligned}
\mathbf{P}\left\{\max_{g\in\mathcal{G}} Z(g) > \sqrt{n}\beta(X_1^{2n}) \middle| (X_i, Y_i)_{i=1}^{2n}\right\} &\leq |\mathcal{G}| e^{-n\beta^2(X_1^{2n})/2} \\
&= e^{-2n\epsilon^2}
\end{aligned}$$

which implies (10).

Now it is a simple matter to to obtain the theorem. In Steps 1 and 2, we set

$$\epsilon(X_1^n) = \sqrt{\left(\frac{18}{n} d(X_1^n) + \epsilon^2\right) \ln\left(\frac{32en}{d(X_1^n)}\right) \ln(128n)},$$

where $n\epsilon^2 \geq 2$ so that $n\epsilon^2(X_1^n) \geq 2$. Also, in Step 5 we set

$$\beta(X_1^n) = \sqrt{\frac{2}{n} \ln \mathcal{N}_\infty(\gamma/2, \mathcal{H}, X_1^{2n}) + 4\epsilon^2}.$$

Then for any $\alpha > 0$, Steps 1-6 imply

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{f\in\mathcal{F}} (L(f) - (1+\alpha)\,\widehat{L}_n^\gamma(f)) > \frac{1+\alpha}{\alpha}\left(\frac{18}{n} d(X_1^n) + \epsilon^2\right) \ln\left(\frac{32en}{d(X_1^n)}\right) \ln(128n)\right\} \\
\leq\ & \mathbf{P}\left\{\sup_{f\in\mathcal{F}} \frac{L(f) - \widehat{L}_n^\gamma(f)}{\sqrt{L(f)}} > \sqrt{\left(\frac{18}{n} d(X_1^n) + \epsilon^2\right) \ln\left(\frac{32en}{d(X_1^n)}\right) \ln(128n)}\right\} \\
\leq\ & 4\mathbf{P}\left\{\sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{(\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f))/2}} > \sqrt{\left(\frac{18}{n} d(X_1^n) + \epsilon^2\right) \ln\left(\frac{32en}{d(X_1^n)}\right) \ln(128n)}\right\} \\
<\ & 4\mathbf{P}\left\{\sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\left(\frac{3}{n} d(X_1^{2n}) + \frac{\epsilon^2}{4}\right) \ln\left(\frac{32en}{d(X_1^{2n})}\right) \ln(128n)}\right\} + 4e^{-\frac{n\epsilon^2}{25}} \\
\leq\ & 4\mathbf{P}\left\{\sup_{f\in\mathcal{F}} \frac{\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)}{\sqrt{\widehat{L'}_n(f) + \widehat{L}_n^\gamma(f)}} > \sqrt{\frac{2}{n} \ln \mathcal{N}_\infty(\gamma/2, \mathcal{H}, X_1^{2n}) + 4\epsilon^2}\right\} + 4e^{-\frac{n\epsilon^2}{25}} \\
\leq\ & 4\mathbf{P}\left\{\max_{g\in\mathcal{G}} \frac{\sum_{i=1}^n \sigma_i (I_{n+i}^{\gamma/2}(g) - I_i^{\gamma/2}(g))}{\sqrt{w(g)}} > \sqrt{2\ln \mathcal{N}_\infty(\gamma/2, \mathcal{H}, X_1^{2n}) + 4n\epsilon^2}\right\} + 4e^{-\frac{n\epsilon^2}{25}} \\
\leq\ & 4e^{-2n\epsilon^2} + 4e^{-\frac{n\epsilon^2}{25}} \\
<\ & 8e^{-\frac{n\epsilon^2}{25}}.
\end{aligned}$$

If $n\epsilon^2 \leq 2$, the same bound obviously holds. Substituting $\epsilon^2 = \frac{25}{n} \ln \frac{8}{\delta}$ yields the theorem. □

Proof of Theorem 1

Step 1 *For any $n \geq 1$ and measurable function $\epsilon(X_1^n)$ of X_1^n such that $n\epsilon^2(X_1^n) \geq 2$ with probability one, we have*

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}(L(f) - \widehat{L}_n^\gamma(f)) > \epsilon(X_1^n)\right\} \leq 4\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\big(\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)\big) > \epsilon(X_1^n)\right\}.$$

Proof. Define $\mathcal{F}'$, a random subset of $\mathcal{F}$, by

$$\mathcal{F}' = \mathcal{F}'(X_1^n) = \{f \in \mathcal{F} : L(f) - \widehat{L}_n^\gamma(f) > \epsilon(X_1^n)\}$$

and note that $I_{\{\mathcal{F}'=\emptyset\}}$ and X_{n+1}^{2n} are independent. As in Step 2 of the previous proof, $nL(f) > n\epsilon^2(X_1^n) \geq 2$ implies that for any $f \in \mathcal{F}'$, $\mathbf{P}\{\widehat{L'}_n(f) > L(f)|X_1^n\} \geq 1/4$, and the argument used there also shows that

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\big(\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)\big) > \epsilon(X_1^n)\right\} \geq \frac{1}{4}\mathbf{P}\{\mathcal{F}' \neq \emptyset\} = \frac{1}{4}\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\big(L(f) - \widehat{L}_n^\gamma(f)\big) > \epsilon(X_1^n)\right\}.$$

Step 2 *For any $\epsilon > 0$,*

$$\begin{aligned}&\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\big(\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)\big) > \sqrt{\left(\frac{9}{n}d(X_1^n) + \frac{\epsilon^2}{2}\right)\ln\left(\frac{32en}{d(X_1^n)}\right)\ln(128n)}\right\}\\ &\leq \quad \mathbf{P}\left\{\sup_{f\in\mathcal{F}}\big(\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)\big) > \sqrt{\frac{2}{n}\ln\mathcal{N}_\infty(\gamma/2, \pi_\gamma(\mathcal{F}), X_1^{2n}) + 4\epsilon^2}\right\} + e^{-\frac{n\epsilon^2}{25}}.\end{aligned}$$

Proof. Combine Steps 3-4 in the proof of Theorem 2.

Step 3 *Suppose $\mathcal{G}$ is a minimal $\gamma/2$-cover of $\pi_\gamma(\mathcal{F})$ and $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables which are also independent of $(X_i, Y_i)_{i=1}^{2n}$. Then for any positive measurable function $\beta(X_1^{2n})$ of X_1^{2n} that depends only on the set $\{X_1, \ldots, X_{2n}\}$,*

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\big(\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)\big) > \beta(X_1^{2n})\right\} \leq \mathbf{P}\left\{\max_{g\in\mathcal{G}}\sum_{i=1}^n \sigma_i\big(I_{n+i}^{\gamma/2}(g) - I_i^{\gamma/2}(g)\big) > n\beta(X_1^{2n})\right\},$$

where $I_i^\gamma(g) = I_{\{\mathrm{margin}(g(X_i),Y_i)<\gamma\}}$.

Proof. As in Step 5 of the proof of Theorem 2, for any $f \in \mathcal{F}$ there is a $g \in \mathcal{G}$ such that $\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f) \leq \widehat{L'}_n^{\gamma/2}(g) - \widehat{L}_n^{\gamma/2}(g)$. Since $\widehat{L'}_n^{\gamma/2}(g) - \widehat{L}_n^{\gamma/2}(g) = \frac{1}{n}\sum_{i=1}^n I_{n+i}^{\gamma/2}(g) - I_i^{\gamma/2}(g)$, the statement follows.

Step 4 *Set*

$$\beta(X_1^n) = \sqrt{\frac{2}{n}\ln\mathcal{N}_\infty(\gamma/2, \mathcal{H}, X_1^{2n}) + 4\epsilon^2},$$

where $\epsilon > 0$. Then

$$\mathbf{P}\left\{\max_{g\in\mathcal{G}}\sum_{i=1}^n \sigma_i(I_{n+i}^{\gamma/2}(g) - I_i^{\gamma/2}(g)) > n\beta(X_1^{2n})\right\} \leq e^{-2n\epsilon^2}.$$

Proof. The claim follows by formally setting $w(g) = n$ in Step 6 of the proof of Theorem 2.

Now in Step 1 set

$$\epsilon(X_1^n) = \sqrt{\left(\frac{9}{n}d(X_1^n) + \frac{\epsilon^2}{2}\right)\ln\left(\frac{32en}{d(X_1^n)}\right)\ln(128n)}.$$

with ϵ is such that $n\epsilon^2 \geq 2$, and in Step 3 set $\beta(X_1^n)$ as in Step 4. Combining Steps 1-4 we obtain

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{f\in\mathcal{F}}(L(f) - \widehat{L}_n^\gamma(f)) > \sqrt{\left(\frac{9}{n}d(X_1^n) + \frac{\epsilon^2}{2}\right)\ln\left(\frac{32en}{d(X_1^n)}\right)\ln(128n)}\right\} \\
&\leq 4\mathbf{P}\left\{\sup_{f\in\mathcal{F}}(\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)) > \sqrt{\left(\frac{9}{n}d(X_1^n) + \frac{\epsilon^2}{2}\right)\ln\left(\frac{32en}{d(X_1^n)}\right)\ln(128n)}\right\} \\
&\leq 4\mathbf{P}\left\{\sup_{f\in\mathcal{F}}(\widehat{L'}_n(f) - \widehat{L}_n^\gamma(f)) > \sqrt{\frac{2}{n}\ln\mathcal{N}_\infty(\gamma/2,\mathcal{H},X_1^{2n}) + 4\epsilon^2}\right\} + 4e^{-\frac{n\epsilon^2}{25}} \\
&\leq 4\mathbf{P}\left\{\max_{g\in\mathcal{G}}\sum_{i=1}^n \sigma_i(I_{n+i}^{\gamma/2}(g) - I_i^{\gamma/2}(g)) > \sqrt{2n\ln\mathcal{N}_\infty(\gamma/2,\mathcal{H},X_1^{2n}) + 4n^2\epsilon^2}\right\} + 4e^{-\frac{n\epsilon^2}{25}} \\
&\leq 4e^{-2n\epsilon^2} + 4e^{-\frac{n\epsilon^2}{25}} \\
&< 8e^{-\frac{n\epsilon^2}{25}}.
\end{aligned}$$

Substituting $\epsilon^2 = \frac{25}{n}\ln\frac{8}{\delta}$ yields Theorem 1. □

Proof of Theorem 3

First note that

$$\begin{aligned}
&\mathbf{E}\sup_{f\in\mathcal{F}}\left(L(f) - \widehat{L}_n^\gamma(f)\right) \\
&= \mathbf{E}\sup_{f\in\mathcal{F}}\mathbf{E}\left[L'_n(f) - \widehat{L}_n^\gamma(f)\middle| D_n\right] \\
&\leq \mathbf{E}\left[\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left(L'_n(f) - \widehat{L}_n^\gamma(f)\right)\middle| D_n\right]\right] \\
&= \mathbf{E}\sup_{f\in\mathcal{F}}\left(L'_n(f) - \widehat{L}_n^\gamma(f)\right) \\
&= \frac{1}{n}\mathbf{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n\left(I_{\{\mathrm{sgn}(f(X_{n+i})-1/2)\neq Y_{n+i}\}} - I_{\{\mathrm{margin}(f(X_i),Y_i)<\gamma\}}\right) \\
&\leq \frac{1}{n}\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n/2}\left(I_{\{\mathrm{sgn}(f(X_{n+i})-1/2)\neq Y_{n+i}\}} - I_{\{\mathrm{margin}(f(X_i),Y_i)<\gamma\}}\right)\right. \\
&\qquad\left. + \sup_{f\in\mathcal{F}}\sum_{i=n/2+1}^{n}\left(I_{\{\mathrm{sgn}(f(X_{n+i})-1/2)\neq Y_{n+i}\}} - I_{\{\mathrm{margin}(f(X_i),Y_i)<\gamma\}}\right)\right]
\end{aligned}$$

$$= \frac{2}{n}\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n/2}\left(I_{\{\mathrm{sgn}(f(X_{n+i})-1/2)\neq Y_{n+i}\}} - I_{\{\mathrm{margin}(f(X_i),Y_i)<\gamma\}}\right)\right]$$
$$= \mathbf{E}\sup_{f\in\mathcal{F}}\left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f)\right) .$$

The proof may be finished by noting that, by McDiarmid's bounded difference inequality (McDiarmid, 1989), for every $\epsilon > 0$,

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left(L(f) - \widehat{L}^{\gamma}_n(f)\right) \geq \mathbf{E}\sup_{f\in\mathcal{F}}\left(L(f) - \widehat{L}^{\gamma}_n(f)\right) + \epsilon\right\} \leq e^{-2n\epsilon^2}$$

and

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f)\right) \leq \mathbf{E}\sup_{f\in\mathcal{F}}\left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f)\right) - \epsilon\right\} \leq e^{-n\epsilon^2/2} .$$

Combining the three inequalities above, we obtain that, for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{\exists f\in\mathcal{F} : L(f) \geq \widehat{L}^{\gamma}_n(f) + \sup_{f\in\mathcal{F}}\left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f)\right) + \epsilon\right\} \\
&\leq \mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left(L(f) - \widehat{L}^{\gamma}_n(f)\right) \geq \sup_{f\in\mathcal{F}}\left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f)\right) + \epsilon\right\} \\
&\leq \mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left(L(f) - \widehat{L}^{\gamma}_n(f)\right) \geq \mathbf{E}\sup_{f\in\mathcal{F}}\left(L(f) - \widehat{L}^{\gamma}_n(f)\right) + \frac{\epsilon}{3}\right\} \\
&\quad + \mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f)\right) \leq \mathbf{E}\sup_{f\in\mathcal{F}}\left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f)\right) - \frac{2\epsilon}{3}\right\} \\
&\leq 2e^{-2n\epsilon^2/9} .
\end{aligned}$$

Setting $\epsilon = \sqrt{(9/(2n))\ln(2/\delta)}$ concludes the proof. □

Proof of Theorem 4

By Theorem 3, with probability at least $1-\delta/2$, for all $f\in\mathcal{F}$ we have

$$L(f) < \widehat{L}^{\gamma}_n(f) + \sup_{f\in\mathcal{F}}\left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f)\right) + 3\sqrt{\frac{\ln(4/\delta)}{2n}}. \tag{11}$$

Defining $\phi_\gamma : \mathbb{R}\to\mathbb{R}$ by

$$\phi_\gamma(t) = \begin{cases} 1 & \text{if } t < 0 \\ 0 & \text{if } t \geq \gamma \\ 1 - t/\gamma & \text{if } t\in[0,\gamma) \end{cases}$$

and noting that for all $t\in\mathbb{R}$, $I_{\{t\leq 0\}} \leq \phi_\gamma(t) \leq I_{\{t<\gamma\}}$, we see that for all $f\in\mathcal{F}$,

$$\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f) \leq \frac{2}{n}\sum_{i=1}^{n/2}\left[\phi_\gamma(\mathrm{margin}(f(X_{n/2+i}),Y_{n/2+i})) - \phi_\gamma(\mathrm{margin}(f(X_i),Y_i))\right] .$$

Using McDiarmid's bounded difference inequality for the supremum of the right-hand side for $f \in \mathcal{F}$, we obtain that with probability at least $1 - \delta/4$,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \left(\widehat{L}^{(2)}_{n/2}(f) - \widehat{L}^{\gamma}_{n/2}(f)\right) \\
&\leq \frac{2}{n}\mathbf{E}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^{n/2}\left[\phi_\gamma(\mathrm{margin}(f(X_{n/2+i}),Y_{n/2+i})) - \phi_\gamma(\mathrm{margin}(f(X_i),Y_i))\right]\right) \\
&\qquad + \sqrt{\frac{2\ln(4/\delta)}{n}} \\
&= \frac{2}{n}\mathbf{E}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^{n/2}\sigma_i\left[\phi_\gamma(\mathrm{margin}(f(X_{n/2+i}),Y_{n/2+i})) - \phi_\gamma(\mathrm{margin}(f(X_i),Y_i))\right]\right) \\
&\qquad + \sqrt{\frac{2\ln(4/\delta)}{n}}
\end{aligned}$$

where $\sigma_1, \ldots, \sigma_{n/2}$ are i.i.d. Rademacher random variables (i.e., $\mathbf{P}(\sigma_1 = 1) = \mathbf{P}(\sigma_1 = -1) = 1/2$) which are also independent of $(X_i, Y_i)_{i=1}^n$. Since ϕ_γ is Lipschitz with parameter $1/\gamma$, we may use the "contraction principle" (see Ledoux and Talagrand, 1991, Theorem 4.2) to obtain

$$\begin{aligned}
&\mathbf{E}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^{n/2}\sigma_i\left[\phi_\gamma(\mathrm{margin}(f(X_{n/2+i}),Y_{n/2+i})) - \phi_\gamma(\mathrm{margin}(f(X_i),Y_i))\right]\right) \\
&\leq \frac{1}{\gamma}\mathbf{E}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^{n/2}\sigma_i\left[\mathrm{margin}(f(X_{n/2+i}),Y_{n/2+i}) - \mathrm{margin}(f(X_i),Y_i)\right]\right) \\
&= \frac{1}{\gamma}\mathbf{E}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^{n/2}\left[\mathrm{margin}(f(X_{n/2+i}),Y_{n/2+i}) - \mathrm{margin}(f(X_i),Y_i)\right]\right) \\
&= \frac{1}{\gamma}\mathbf{E}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^{n/2}\left[(f(X_{n/2+i}) - 1/2)\tilde{Y}_{n/2+i} - (f(X_i) - 1/2)\tilde{Y}_i\right]\right) \\
&\qquad \text{where } \tilde{Y}_i = 2Y_i - 1 \\
&= \frac{1}{\gamma}\mathbf{E}\left(\sup_{\substack{N,w_1,\ldots,w_N\\ h_1,\ldots,h_N\in\mathcal{H}}}\sum_{j=1}^{N} w_j\sum_{i=1}^{n/2}\left[(h_j(X_{n/2+i}) - 1/2)\tilde{Y}_{n/2+i} - (h_j(X_i) - 1/2)\tilde{Y}_i\right]\right) \\
&= \frac{1}{\gamma}\mathbf{E}\left(\sup_{h\in\mathcal{H}}\sum_{i=1}^{n/2}\left[(h(X_{n/2+i}) - 1/2)\tilde{Y}_{n/2+i} - (h(X_i) - 1/2)\tilde{Y}_i\right]\right)
\end{aligned}$$

where the last equality follows from the fact that the supremum of the previous line is always achieved by a convex combination concentrated on just one base classifier. Once again, using the bounded difference inequality, we note that with probability at least $1-\delta/4$,

$$\begin{aligned}
&\mathbf{E}\left(\sup_{h\in\mathcal{H}}\sum_{i=1}^{n/2}\left[(h(X_{n/2+i})-1/2)\tilde{Y}_{n/2+i}-(h(X_i)-1/2)\tilde{Y}_i\right]\right)\\
&\le \sup_{h\in\mathcal{H}}\sum_{i=1}^{n/2}\left[(h(X_{n/2+i})-1/2)\tilde{Y}_{n/2+i}-(h(X_i)-1/2)\tilde{Y}_i\right]+\sqrt{\frac{n\ln(4/\delta)}{2}}\\
&= \frac{n}{2}\sup_{h\in\mathcal{H}}\left(\widehat{L}^{(1)}_{n/2}(h)-\widehat{L}^{(2)}_{n/2}(h)\right)+\sqrt{\frac{n\ln(4/\delta)}{2}}.
\end{aligned}$$

The statement of the theorem now follows by putting the pieces together using the union bound. □

Proof of Lemma 5

First we show that if (4) holds, $\mathcal{F}_\alpha$ γ-shatters x_1^n. Let $(b_1,\ldots,b_n)\in\{0,1\}^n$ be an arbitrary binary vector, and let $\tilde{b}_i=2b_i-1$ for $i=1,\ldots,n$. For $j=2,\ldots,n$ we define $w_j=\frac{\frac{1}{d_j}}{\sum_{i=2}^n\frac{1}{d_i}}$ and $g_j=g^{(x_{j-1},x_j,-\tilde{b}_j\alpha d_j+r,-\tilde{b}_{j-1}\alpha d_j+r)}$ where $r=\frac{\alpha}{\sum_{j=2}^n\frac{1}{d_j}}\sum_{i=1}^n\tilde{b}_i$. By the definitions of $\mathcal{G}_\alpha$ and d_i it is clear that $g_j\in\mathcal{G}_\alpha$ for $j=2,\ldots,n$. Since $\sum_{j=2}^n w_j=1$, $f(x)=\sum_{j=2}^n w_jg_j(x)\in\mathcal{F}_\alpha$.

We show by induction that $\tilde{b}_if(x_i)=\frac{\alpha}{\sum_{j=2}^n\frac{1}{d_j}}$ for all $i=1,\ldots,n$. This together with (4) means that $\tilde{b}_if(x_i)\ge\gamma$ for all $i=1,\ldots,n$; hence (1) is satisfied with $r_i=0$ for all $i=1,\ldots,n$. For x_1 we have

$$\begin{aligned}
\tilde{b}_1f(x_1) &= \tilde{b}_1\sum_{i=2}^n w_ig_i(x_1)\\
&= \tilde{b}_1\sum_{i=2}^n\frac{\frac{1}{d_i}}{\sum_{j=2}^n\frac{1}{d_j}}(-\tilde{b}_i\alpha d_i+r)\\
&= -\tilde{b}_1\frac{\alpha}{\sum_{j=2}^n\frac{1}{d_j}}\sum_{i=2}^n\tilde{b}_i+\tilde{b}_1r\frac{1}{\sum_{j=2}^n\frac{1}{d_j}}\sum_{i=2}^n\frac{1}{d_i}\\
&= -\tilde{b}_1\frac{\alpha}{\sum_{j=2}^n\frac{1}{d_j}}\sum_{i=2}^n\tilde{b}_i+\tilde{b}_1r\\
&= -\tilde{b}_1\frac{\alpha}{\sum_{j=2}^n\frac{1}{d_j}}\sum_{i=2}^n\tilde{b}_i+\tilde{b}_1\frac{\alpha}{\sum_{j=2}^n\frac{1}{d_j}}\sum_{i=1}^n\tilde{b}_i\\
&= \frac{\alpha}{\sum_{j=2}^n\frac{1}{d_j}}.
\end{aligned}$$

In the inductive step we assume that $\tilde{b}_{i-1}f(x_{i-1})=\frac{\alpha}{\sum_{j=2}^n\frac{1}{d_j}}$. Since the only base function that can change between x_{i-1} and x_i is g_i, we have

$$\tilde{b}_if(x_i) = \tilde{b}_if(x_{i-1})+\tilde{b}_iw_i\left(g_i(x_i)-g_i(x_{i-1})\right)$$

$$
\begin{aligned}
&= \tilde{b}_i f(x_{i-1}) + \tilde{b}_i w_i \left(-\tilde{b}_{i-1} \alpha d_i + \tilde{b}_i \alpha d_i \right) \\
&= \tilde{b}_i f(x_{i-1}) + w_i \alpha d_i (1 - \tilde{b}_i \tilde{b}_{i-1}).
\end{aligned}
$$

If $\tilde{b}_{i-1} = \tilde{b}_i$ then $\tilde{b}_i f(x_i) = \tilde{b}_i f(x_{i-1}) = \tilde{b}_{i-1} f(x_{i-1}) = \frac{\alpha}{\sum_{j=2}^n \frac{1}{d_j}}$. If $\tilde{b}_{i-1} \neq \tilde{b}_i$ then

$$
\begin{aligned}
\tilde{b}_i f(x_i) &= -\tilde{b}_{i-1} f(x_{i-1}) + 2 w_i \alpha d_i \\
&= -\frac{\alpha}{\sum_{j=2}^n \frac{1}{d_j}} + 2 \frac{\alpha}{\sum_{j=2}^n \frac{1}{d_j}} \\
&= \frac{\alpha}{\sum_{j=2}^n \frac{1}{d_j}}.
\end{aligned}
$$

Next we show that if $\mathcal{F}_\alpha$ γ-shatters x_1^n, (4) holds. Consider the two alternating labeling $\tilde{b}_i^+ = -\tilde{b}_i^- = (-1)^i, i = 1, \ldots, n$. If $\mathcal{F}_\alpha$ γ-shatters x_1^n then by (1), there exists $f^+, f^- \in \mathcal{F}_\alpha$ such that for a given real vector $(r_1, \ldots, r_n)$,

$$
\begin{aligned}
\tilde{b}_i^+ f^+(x_i) &\geq \tilde{b}_i^+ r_i + \gamma, \\
\tilde{b}_i^- f^-(x_i) &\geq \tilde{b}_i^- r_i + \gamma,
\end{aligned}
$$

for $i = 1, \ldots, n$, so by setting $f(x) = \frac{1}{2}(f^+(x) - f^-(x))$,

$$
(-1)^i f(x_i) \geq \gamma.
$$

By the definition of $\mathcal{G}_\alpha$, if $g \in \mathcal{G}_\alpha$ then $-g \in \mathcal{G}_\alpha$, so $f \in \mathcal{F}_\alpha$, which means that f can be written in the form

$$
f = \sum_{j=1}^N w_j g^{(x_{a_j}, x_{b_j}, y_{a_j}, y_{b_j})}
$$

where $\sum_{j=1}^N w_j = 1$. Let $\alpha_j = \frac{1}{2} \left| \frac{y_b - y_a}{x_b - x_a} \right|$ and $s_j = \operatorname{sgn}\left(\frac{y_b - y_a}{x_b - x_a} \right)$ for $j = 1, \ldots, N$. Since $(-1)^i (f(x_i) - f(x_{i-1})) \geq 2\gamma$ for all $i = 2, \ldots, n$, and since f is continuous, there must be a point x_i' between x_{i-1} and x_i where $(-1)^i$ times the (left) derivative of f is not less then $\frac{2\gamma}{x_i - x_{i-1}}$. Therefore,

$$
(-1)^i f'(x_i') = (-1)^i \sum_{j : x_{a_j} < x_i' \leq x_{b_j}} w_j 2 s_j \alpha_j \geq \frac{2\gamma}{d_i}
$$

for $i = 2, \ldots, n$. Taking the sum of both sides from $i = 2$ to n yields

$$
\begin{aligned}
\gamma \sum_{i=2}^n \frac{1}{d_i} &\leq \sum_{i=2}^n (-1)^i \sum_{j : x_{a_j} < x_i' \leq x_{b_j}} w_j s_j \alpha_j \\
&= \sum_{j=1}^N w_j s_j \alpha_j \sum_{i : x_{a_j} < x_i' \leq x_{b_j}} (-1)^i
\end{aligned}
$$

$$
\begin{aligned}
&\le \sum_{j=1}^{N} w_j \alpha_j \qquad (\text{since } x'_2 < \ldots < x'_n) \\
&\le \alpha.
\end{aligned}
$$

□

Acknowledgments

We thank Miklós Csűrös for helpful comments and discussions. We are grateful to Vladimir Koltchinskii for pointing out a key error in an earlier version of this paper. We also thank the reviewers for valuable suggestions. This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and DGI grant BMF2000-0807. A preliminary version of the paper was presented at The Fourteenth Annual Conference on Computational Learning Theory, Amsterdam, July 16–19, 2001

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, Cambridge, 1999.

P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.

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

P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, to appear, 2002.

C. Blake, E. Keogh, and C. Merz. Uci repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`, 1998.

S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York, 2000.

T. Erlebach, K. Jansen, and E. Seidel. Polynomial-time approximation schemes for geometric graphs. In *Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms SODA'01*, pages 671–679, 2001.

J. Hastad. Clicque is hard to approximate within $n^{1-\epsilon}$. In *Proceedings of the 37th Annual Symposium on Foundations of Computer Science FOCS'96*, pages 627–636, 1996.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

M. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer Systems Sciences*, 48:464–497, 1994.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44:1926–1940, 1998.

J. Shawe-Taylor and R.C. Williamson. Generalization performance of classifiers in terms of observed covering numbers. In H. U. Simon P. Fischer, editor, *Computational Learning Theory: Proceedings of the Fourth European Conference, EuroCOLT'99*, pages 153–167. Springer, Berlin, 1999. Lecture Notes in Artificial Intelligence 1572.

E.V. Slud. Distribution inequalities for the binomial law. *Annals of Probability*, 5:404–412, 1977.

V.N. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.