Journal of Machine Learning Research v (2002) pages Submitted 11/01; Published 10/02

Rademacher and Gaussian Complexities: Risk Bounds and Structural Results

Peter L. Bartlett PETER.BARTLETT@ANU.EDU.AU
Shahar Mendelson SHAHAR@CSL.ANU.EDU.AU
Research School of Information Sciences and Engineering
Australian National University
Canberra 0200, Australia

Editor: Philip M. Long

Abstract

We investigate the use of certain data-dependent estimates of the complexity of a function class, called Rademacher and gaussian complexities. In a decision theoretic setting, we prove general risk bounds in terms of these complexities. We consider function classes that can be expressed as combinations of functions from basis classes and show how the Rademacher and gaussian complexities of such a function class can be bounded in terms of the complexity of the basis classes. We give examples of the application of these techniques in finding data-dependent risk bounds for decision trees, neural networks and support vector machines.

Keywords: Error Bounds, Data-Dependent Complexity, Rademacher Averages, Maximum Discrepancy

1. Introduction

In learning problems like pattern classification and regression, a considerable amount of effort has been spent on obtaining good error bounds. These are useful, for example, for the problem of model selection—choosing a model of suitable complexity. Typically, such bounds take the form of a sum of two terms: some sample-based estimate of performance and a penalty term that is large for more complex models. For example, in pattern classification, the following theorem is an improvement of a classical result of Vapnik and Chervonenkis (Vapnik and Chervonenkis, 1971).

Theorem 1 *Let F be a class of $\{\pm 1\}$-valued functions defined on a set $\mathcal{X}$. Let P be a probability distribution on $\mathcal{X} \times \{\pm 1\}$, and suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ and (X, Y) are chosen independently according to P. Then, there is an absolute constant c such that for any integer n, with probability at least $1 - \delta$ over samples of length n, every f in F satisfies*

$$P(Y \neq f(X)) \leq \hat{P}_n(Y \neq f(X)) + c\sqrt{\frac{\mathrm{VCdim}(F)}{n}},$$

where $\mathrm{VCdim}(F)$ *denotes the Vapnik-Chervonenkis dimension of F,*

$$\hat{P}_n(S) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_S(X_i, Y_i),$$

and $\mathbf{1}_S$ *is the indicator function of* S.

In this case, the sample-based estimate of performance is the proportion of examples in the training sample that are misclassified by the function f, and the complexity penalty term involves the VC-dimension of the class of functions. It is natural to use such bounds for the model selection scheme known as complexity regularization: choose the model class containing the function with the best upper bound on its error. The performance of such a model selection scheme critically depends on how well the error bounds match the true error (see Bartlett et al., 2002). There is theoretical and experiment evidence that error bounds involving a fixed complexity penalty (that is, a penalty that does not depend on the training data) cannot be universally effective (Kearns et al., 1997).

Recently, several authors have considered alternative notions of the complexity of a function class: the maximum discrepancy (Bartlett et al., 2002) and the Rademacher and gaussian complexities (see Bartlett et al., 2002, Koltchinskii, 2000, Koltchinskii and Panchenko, 2000a,b, Mendelson, 2001b).

Definition 2 *Let* μ *be a probability distribution on a set* $\mathcal{X}$ *and suppose that* $X_1, \ldots, X_n$ *are independent samples selected according to* μ. *Let* F *be a class of functions mapping from* $\mathcal{X}$ *to* $\mathbb{R}$. *Define the* maximum discrepancy *of* F *as the random variable*

$$\hat{D}_n(F) = \sup_{f \in F} \left(\frac{2}{n} \sum_{i=1}^{n/2} f(X_i) - \frac{2}{n} \sum_{i=n/2+1}^{n} f(X_i) \right).$$

Denote the expected maximum discrepancy *of* F *by* $D_n(F) = \mathbf{E}\hat{D}_n(F)$.

Define the random variable

$$\hat{R}_n(F) = \mathbf{E}\left[\sup_{f \in F} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right| \middle| X_1, \ldots, X_n \right],$$

where $\sigma_1, \ldots, \sigma_n$ *are independent uniform* $\{\pm 1\}$*-valued random variables. Then the* Rademacher complexity *of* F *is* $R_n(F) = \mathbf{E}\hat{R}_n(F)$. *Similarly, define the random variable*

$$\hat{G}_n(F) = \mathbf{E}\left[\sup_{f \in F} \left| \frac{2}{n} \sum_{i=1}^{n} g_i f(X_i) \right| \middle| X_1, \ldots, X_n \right],$$

where $g_1, \ldots, g_n$ *are independent gaussian* $N(0,1)$ *random variables. The* gaussian complexity *of* F *is* $G_n(F) = \mathbf{E}\hat{G}_n(F)$.

All three quantities are intuitively reasonable as measures of complexity of the function class F: $\hat{D}_n(F)$ quantifies how much the behavior on half of the sample can be unrepresentative of the behavior on the other half, and both $R_n(F)$ and $G_n(F)$ quantify the extent to which some function in the class F can be correlated with a noise sequence of length n. The following two lemmas show that these complexity measures are closely related. The proof of the first is in Appendix A; the second is from (Tomczak-Jaegermann, 1989).

Lemma 3 *Let F be a class of functions that map to $[-1,1]$. Then for every integer n,*

$$\frac{R_n(F)}{2} - 2\sqrt{\frac{2}{n}} \leq D_n(F) \leq R_n(F) + 4\sqrt{\frac{2}{n}}.$$

If F is closed under negation, the lower bound can be strengthened to

$$R_n(F) - 4\sqrt{\frac{2}{n}} \leq D_n(F).$$

Furthermore,

$$P\left\{\left|\hat{D}_n(F) - D_n(F)\right| \geq \epsilon\right\} \leq 2\exp\left(\frac{-\epsilon^2 n}{2}\right).$$

Lemma 4 *There are absolute constants c and C such that for every class F and every integer n, $cR_n(F) \leq G_n(F) \leq C \ln n R_n(F)$.*

The following theorem is an example of the usefulness of these notions of complexity. The proof of the first part is in (Bartlett et al., 2002). The proof of the second part is a slight refinement of a proof of a more general result which we give below (Theorem 8); it is presented in Appendix B.

Theorem 5 *Let P be a probability distribution on $\mathcal{X} \times \{\pm 1\}$, let F be a set of $\{\pm 1\}$-valued functions defined on $\mathcal{X}$, and let $(X_i, Y_i)_{i=1}^n$ be training samples drawn according to P^n.*

(a) With probability at least $1-\delta$, every function f in F satisfies

$$P(Y \neq f(X)) \leq \hat{P}_n(Y \neq f(X)) + \hat{D}_n(F) + \sqrt{\frac{9\ln(1/\delta)}{2n}}\,.$$

(b) With probability at least $1-\delta$, every function f in F satisfies

$$P(Y \neq f(X)) \leq \hat{P}_n(Y \neq f(X)) + \frac{R_n(F)}{2} + \sqrt{\frac{\ln(1/\delta)}{2n}}\,.$$

The following result shows that this theorem implies the upper bound of Theorem 1 in terms of VC-dimension, as well as a refinement in terms of VC-entropy. In particular, Theorem 5 can never be much worse than the VC results. Since the proof of Theorem 5 is a close analog of the first step of the proof of VC-style results (the symmetrization step), this is not surprising. In fact, the bounds of Theorem 5 can be considerably better than Theorem 1, since the first part of the following result is in terms of the *empirical* VC-dimension.

Theorem 6 *Fix a sample $X_1, \ldots, X_n$. For a function class $F \subseteq \{\pm 1\}^{\mathcal{X}}$, define the restriction of F to the sample as*

$$F_{|X_i} = \{(f(X_1), \ldots, f(X_n)) : f \in F\}\,.$$

Define the empirical VC-dimension of F as $d = \mathrm{VCdim}\left(F_{|X_i}\right)$ and the empirical VC-entropy of F as $E = \log_2\left|F_{|X_i}\right|$. Then $\hat{G}_n(F) = O\left(\sqrt{d/n}\right)$ and $\hat{G}_n(F) = O\left(\sqrt{E/n}\right)$.

The proof of this theorem is based on an upper bound on $\hat{G}_n$ which is due to Dudley, together with an upper bound on covering numbers due to Haussler (see Mendelson, 2001a).

Koltchinskii and Panchenko (2000a) proved an analogous error bound in terms of *margins*. The margin of a real-valued function f on a labelled example $(x, y) \in \mathcal{X} \times \{\pm 1\}$ is $yf(x)$. For a function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, we write

$$\hat{\mathbf{E}}_n h(X, Y) = (1/n) \sum_{i=1}^{n} h(X_i, Y_i).$$

Theorem 7 *Let P be a probability distribution on $\mathcal{X} \times \{\pm 1\}$ and let F be a set of real-valued functions defined on $\mathcal{X}$, with $\sup\{|f(x)| : f \in F\}$ finite for all $x \in \mathcal{X}$. Suppose that $\phi : \mathbb{R} \to [0, 1]$ satisfies $\phi(\alpha) \geq \mathbf{1}(\alpha \leq 0)$ and is Lipschitz with constant L. Then with probability at least $1 - \delta$ with respect to training samples $(X_i, Y_i)_{i=1}^n$ drawn according to P^n, every function in F satisfies*

$$P(Yf(X) \leq 0) \leq \hat{\mathbf{E}}_n \phi(Yf(X)) + 2LR_n(F) + \sqrt{\frac{\ln(2/\delta)}{2n}}\,.$$

This improves a number of results bounding error in terms of a sample average of a margin error plus a penalty term involving the complexity of the real-valued class (such as covering numbers and fat-shattering dimensions; see Bartlett, 1998, Mason et al., 2000, Schapire et al., 1998, Shawe-Taylor et al., 1998).

In the next section, we give a bound of this form that is applicable in a more general, decision-theoretic setting. Here, we have an input space $\mathcal{X}$, an action space $\mathcal{A}$ and an output space $\mathcal{Y}$. The n training examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are selected independently according to a probability measure P on $\mathcal{X} \times \mathcal{Y}$. There is a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{A} \to [0, 1]$, so that $\mathcal{L}(y, a)$ reflects the cost of taking a particular action $a \in \mathcal{A}$ when the outcome is $y \in \mathcal{Y}$. The aim of learning is to choose a function f that maps from $\mathcal{X}$ to $\mathcal{A}$, so as to minimize the expected loss $\mathbf{E}\mathcal{L}(Y, f(X))$.

For example, in multiclass classification, the output space $\mathcal{Y}$ is the space $\mathcal{Y} = \{1, \ldots, k\}$ of class labels. When using error correcting output codes (Kong and Dietterich, 1995, Schapire, 1997) for this problem, the action space might be $\mathcal{A} = [0, 1]^m$, and for each $y \in \mathcal{Y}$ there is a codeword $a_y \in \mathcal{A}$. The loss function $\mathcal{L}(y, a)$ is equal to 0 if the closest codeword a_{y^*} has $y^* = y$ and 1 otherwise.

Section 2 gives bounds on the expected loss for decision-theoretic problems of this kind in terms of the sample average of a Lipschitz *dominating cost function* (a function that is pointwise larger than the loss function) plus a complexity penalty term involving a Rademacher complexity.

We also consider the problem of estimating $R_n(F)$ and $G_n(F)$ (for instance, for model selection). These quantities can be estimated by solving an optimization problem over F. However, for cases of practical interest, such optimization problems are difficult. On the other hand, in many such cases, functions in F can be represented as combinations of functions from simpler classes. This is the case, for instance, for decision trees, voting methods, and neural networks. In Section 3, we show how the complexity of such a class can be related to the complexity of the class of basis functions. Section 4 describes examples of the application of these techniques.

An earlier version of this paper appeared in COLT'01 (Bartlett and Mendelson, 2001).

2. Risk Bounds

We begin with some notation. Given an independent sample $(X_i, Y_i)_{i=1}^n$ distributed as (X, Y), we denote by P_n the empirical measure supported on that sample and by μ_n the empirical measure supported on $(X_i)_{i=1}^n$. We say a function $\phi : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ dominates a loss function $\mathcal{L}$ if for all $y \in \mathcal{Y}$ and $a \in \mathcal{A}$, $\phi(y, a) \geq \mathcal{L}(y, a)$. For a class of functions F, conv$F$ is the class of convex combinations of functions from F, $-F = \{-f : f \in F\}$, absconv$F$ is the class of convex combinations of functions from $F \cup -F$, and $cF = \{cf : f \in F\}$. If ϕ is a function defined on the range of the functions in F, let $\phi \circ F = \{\phi \circ f | f \in F\}$. Given a set A, we denote its characteristic function by $\mathbf{1}_A$ or $\mathbf{1}(A)$. Finally, constants are denoted by C or c. Their values may change from line to line, or even within the same line.

Theorem 8 *Consider a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{A} \to [0, 1]$ and a dominating cost function $\phi : \mathcal{Y} \times \mathcal{A} \to [0, 1]$. Let F be a class of functions mapping from $\mathcal{X}$ to $\mathcal{A}$ and let $(X_i, Y_i)_{i=1}^n$ be independently selected according to the probability measure P. Then, for any integer n and any $0 < \delta < 1$, with probability at least $1 - \delta$ over samples of length n, every f in F satisfies*

$$\mathbf{E}\mathcal{L}(Y, f(X)) \leq \hat{\mathbf{E}}_n \phi(Y, f(X)) + R_n(\tilde{\phi} \circ F) + \sqrt{\frac{8 \ln(2/\delta)}{n}} \, ,$$

where $\tilde{\phi} \circ F = \{(x, y) \mapsto \phi(y, f(x)) - \phi(y, 0) : f \in F\}$.

The proof uses McDiarmid's inequality (McDiarmid, 1989).

Theorem 9 (McDiarmid's Inequality) *Let $X_1, ..., X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1, ..., x_n, x_i' \in A} \left| f(x_1, ..., x_n) - f(x_1, ..., x_{i-1}, x_i', x_{i+1}, ...x_n) \right| \leq c_i$$

for every $1 \leq i \leq n$. Then, for every $t > 0$,

$$P\left\{ f(X_1, ..., X_n) - \mathbf{E} f(X_1, ..., X_n) \geq t \right\} \leq e^{-2t^2 / \sum_{i=1}^n c_i^2}.$$

Proof *(of Theorem 8)* Since ϕ dominates $\mathcal{L}$, for all $f \in F$ we can write

$$\begin{aligned}
\mathbf{E}\mathcal{L}(Y, f(X)) &\leq \mathbf{E}\phi(Y, f(X)) \\
&\leq \hat{\mathbf{E}}_n \phi(Y, f(X)) + \sup_{h \in \phi \circ F} \left(\mathbf{E}h - \hat{\mathbf{E}}_n h \right) \\
&= \hat{\mathbf{E}}_n \phi(Y, f(X)) + \sup_{h \in \tilde{\phi} \circ F} \left(\mathbf{E}h - \hat{\mathbf{E}}_n h \right) \\
&\qquad + \mathbf{E}\phi(Y, 0) - \hat{\mathbf{E}}_n \phi(Y, 0).
\end{aligned}$$

When an (X_i, Y_i) pair changes, the random variable $\sup_{h \in \tilde{\phi} \circ F} \left(\mathbf{E}h - \hat{\mathbf{E}}_n h \right)$ can change by no more than $2/n$. McDiarmid's inequality implies that with probability at least $1 - \delta/2$,

$$\sup \left(\mathbf{E}h - \hat{\mathbf{E}}_n h \right) \leq \mathbf{E} \sup \left(\mathbf{E}h - \hat{\mathbf{E}}_n h \right) + \sqrt{2 \ln(2/\delta)/n} \, .$$

A similar argument, together with the fact that $\mathbf{E}\hat{\mathbf{E}}_n\phi(Y,0) = \mathbf{E}\phi(Y,0)$, shows that with probability at least $1-\delta$,

$$\mathbf{E}\mathcal{L}(Y,f(X)) \le \hat{\mathbf{E}}_n\phi(Y,f(X)) + \mathbf{E}\sup_{h\in\tilde{\phi}\circ F}\left(\mathbf{E}h - \hat{\mathbf{E}}_n h\right) + \sqrt{\frac{8\ln(2/\delta)}{n}}.$$

It remains to show that the second term on the right hand side is no more than $R_n(\tilde{\phi}\circ F)$. If $(X_1',Y_1'),\ldots,(X_n',Y_n')$ are independent random variables with the same distribution as (X,Y), then

$$\begin{aligned}
\mathbf{E}\sup_{h\in\tilde{\phi}\circ F}\left(\mathbf{E}h - \hat{\mathbf{E}}_n h\right) &= \mathbf{E}\sup_{h\in\tilde{\phi}\circ F}\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n h(X_i',Y_i') - \hat{\mathbf{E}}_n h \,\middle|\, (X_i,Y_i)\right] \\
&\le \mathbf{E}\sup_{h\in\tilde{\phi}\circ F}\left(\frac{1}{n}\sum_{i=1}^n h(X_i',Y_i') - \hat{\mathbf{E}}_n h\right) \\
&= \mathbf{E}\sup_{h\in\tilde{\phi}\circ F}\frac{1}{n}\sum_{i=1}^n \sigma_i\left(h(X_i',Y_i') - h(X_i,Y_i)\right) \\
&\le 2\mathbf{E}\sup_{h\in\tilde{\phi}\circ F}\frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i,Y_i) \\
&\le R_n(\tilde{\phi}\circ F).
\end{aligned}$$

■

As an example, consider the case $\mathcal{A} = \mathcal{Y} = [0,1]$. It is possible to bound $R_n(F)$ in terms of expected covering numbers of F or its fat-shattering dimension. Indeed, the following result relates $G_n(F)$ to empirical versions of these notions of complexity, and implies that Theorem 8 can never give a significantly worse estimate than previous estimates in terms of these quantities. This result is essentially in (Mendelson, 2001b); although that paper gave a result in terms of the fat-shattering dimension, the same proof works for the empirical fat-shattering dimension.

Theorem 10 *Fix a sample $X_1,\ldots,X_n$. Let F be a class of functions whose range is contained in $[-1,1]$. Assume that there is some $\gamma > 1$ such that for any $\epsilon > 0$, $\mathrm{fat}_\epsilon\left(F_{|X_i}\right) \le \gamma\epsilon^{-p}$. Then, there are absolute constants C_p, which depend only on p, such that*

$$\hat{G}_n(F) \le \begin{cases} C_p\gamma^{1/2}\ln\gamma\, n^{-1/2} & \text{if } 0<p<2, \\ C_2\left(\gamma^{1/2}\ln\gamma\right)n^{-1/2}\ln^2 n & \text{if } p=2, \\ C_p\left(\gamma^{1/2}\ln\gamma\right)n^{-1/p} & \text{if } p>2. \end{cases}$$

3. Estimating the Rademacher and Gaussian Complexities of Function Classes

An important property of Rademacher complexity is that it can be estimated from a single sample $(X_1,\ldots,X_n)$, and from a single realization of the Rademacher variables. The following result follows from McDiarmid's inequality. A similar result is true for the gaussian complexity.

Theorem 11 *Let F be a class of functions mapping to $[-1,1]$. For any integer n,*

$$P\left\{\left|R_n(F) - \frac{2}{n}\sup_{f\in F}\left|\sum_{i=1}^{n}\sigma_i f(X_i)\right|\right| \geq \epsilon\right\} \leq 2\exp\left(\frac{-\epsilon^2 n}{8}\right) ,$$

and

$$P\left\{\left|R_n(F) - \hat{R}_n(F)\right| \geq \epsilon\right\} \leq 2\exp\left(\frac{-\epsilon^2 n}{8}\right) ,$$

Thus, it seems that estimation of $R_n(F)$ and $G_n(F)$ is particularly convenient. However, as mentioned before, the computation involves an optimization over the class F, which is hard for interesting function classes. The way we bypass this obstacle is to use that fact that some "large" classes can be expressed as combinations of functions from simpler classes. For instance, a decision tree can be expressed as a fixed boolean function of the functions appearing in each decision node, voting methods use thresholded convex combinations of functions from a simpler class, and neural networks are compositions of fixed squashing functions with linear combinations of functions from some class. Hence, we present several structural results that lead to bounds on the Rademacher and gaussian complexities of a function class F in terms of the complexities of simpler classes of functions from which F is constructed.

3.1 Simple Structural Results

We begin with the following observations regarding $R_n(F)$.

Theorem 12 *Let $F, F_1, \ldots, F_k$ and H be classes of real functions. Then*

1. *If $F \subseteq H$, $R_n(F) \leq R_n(H)$.*
2. *$R_n(F) = R_n(\mathrm{conv}F) = R_n(\mathrm{absconv}F)$.*
3. *For every $c \in \mathbb{R}$, $R_n(cF) = |c|R_n(F)$.*
4. *If $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L_ϕ and satisfies $\phi(0) = 0$, then $R_n(\phi \circ F) \leq 2L_\phi R_n(F)$.*
5. *For any uniformly bounded function h, $R_n(F + h) \leq R_n(F) + \|h\|_\infty/\sqrt{n}$.*
6. *For $1 \leq q < \infty$, let $\mathcal{L}_{F,h,q} = \{|f - h|^q \,|f \in F\}$, where h is uniformly bounded. If $\|f - h\|_\infty \leq 1$ for every $f \in F$, then $R_n(\mathcal{L}_{F,h,q}) \leq 2q\left(R_n(F) + \|h\|_\infty/\sqrt{n}\right)$.*
7. *$R_n\left(\sum_{i=1}^k F_i\right) \leq \sum_{i=1}^k R_n(F_i)$.*

Parts 1-3 are true for G_n, with exactly the same proof. The other observations hold for G_n with an additional factor of $\ln n$ and may be established using the general connection between R_n and G_n (Lemma 4). Parts 5 and 6 allow us to estimate the Rademacher complexities of natural loss function classes.

Note that 7 is tight. To see this, let $F_1 = \cdots = F_k = F$. Then, by parts 1 and 3, $R_n\left(\sum_{i=1}^k F_i\right) \geq R_n(kF) = kR_n(F) = \sum_{i=1}^k R_n(F_i)$.

Proof Parts 1 and 3 are immediate from the definitions. To see part 2, notice that for every $x_1, \ldots, x_n$ and $\sigma_1, \ldots, \sigma_n$,

$$\begin{aligned}
&\sup_{f \in \mathrm{absconv} F} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \\
&= \sup_{f \in \mathrm{conv} F} \left| \sum \sigma_i f(x_i) \right| \\
&= \max \left(\sup_{f \in \mathrm{conv} F} \sum \sigma_i f(x_i), \sup_{f \in \mathrm{conv} F} - \sum \sigma_i f(x_i) \right) \\
&= \max \left(\sup_{f \in F} \sum \sigma_i f(x_i), \sup_{f \in F} - \sum \sigma_i f(x_i) \right) \\
&= \sup_{f \in F} \left| \sum \sigma_i f(x_i) \right| .
\end{aligned}$$

The inequality of part 4 is due to Ledoux and Talagrand (1991, Corollary 3.17). As for part 5, note that for every realization of $X_1, ..., X_n$,

$$\begin{aligned}
&\mathbf{E} \sup_{f \in F} \left| \sum_{i=1}^n \sigma_i \left(f(x_i) + h(x_i) \right) \right| \\
&\leq \mathbf{E} \sup_{f \in F} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| + \mathbf{E} \left| \sum_{i=1}^n \sigma_i h(x_i) \right| \\
&\leq \mathbf{E} \sup_{f \in F} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| + \left(\sum_{i=1}^n h(x_i)^2 \right)^{\frac{1}{2}} ,
\end{aligned}$$

where the last inequality follows since for any function g, $\mathbf{E}|g| \leq (\mathbf{E}g^2)^{1/2}$. Hence,

$$R_n(F + h) \leq R_n(F) + \|h\|_\infty / \sqrt{n},$$

as claimed.

To see part 6, notice that $\phi(x) = |x|^q$ is a Lipschitz function which passes through the origin with a Lipschitz constant q. By parts 4 and 5 of Theorem 12,

$$R_n(\mathcal{L}_{F,h,q}) \leq 2q R_n(F - h) \leq 2q \left(R_n(F) + \frac{\|h\|_\infty}{\sqrt{n}} \right).$$

Finally, part 7 follows from the triangle inequality. ■

3.2 Lipschitz Functions on $\mathbb{R}^k$

Theorem 12 part 4 shows that composing real-valued functions in some class with a Lipschitz function changes the Rademacher complexity by no more than a constant factor. In this section, we prove a similar result for the gaussian complexity of a class of vector-valued functions.

We require the following comparison theorem for gaussian processes which is due to Slepian (Pisier, 1989).

Lemma 13 *Let $\{X_i, 1 \leq i \leq m\}$ and $\{Y_i, 1 \leq i \leq m\}$ be two gaussian processes which satisfy that, for every i, j,*

$$\|X_i - X_j\|_2 \leq \|Y_i - Y_j\|_2,$$

where $\|X_i - X_j\|_2^2 = \mathbf{E}(X_i - X_j)^2$. Then

$$\mathbf{E} \sup_i X_i \leq 2\mathbf{E} \sup_i Y_i.$$

Now, we can formulate and prove the main result of this section, in which we estimate the gaussian averages of a Lipschitz image of a direct sum of classes.

Theorem 14 *Let $\mathcal{A} = \mathbb{R}^m$ and let F be a class of functions mapping from $\mathcal{X}$ to $\mathcal{A}$. Suppose that there are real-valued classes $F_1, \ldots, F_m$ such that F is a subset of their direct sum. Assume further that $\phi : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ is such that, for all $y \in \mathcal{Y}$, $\phi(y, \cdot)$ is a Lipschitz function (with respect to euclidean distance on $\mathcal{A}$) with constant L which passes through the origin and is uniformly bounded. For $f \in F$, define $\phi \circ f$ as the mapping $(x, y) \mapsto \phi(y, f(x))$. Then, for every integer n and every sample $(X_1, Y_1), \ldots, (X_n, Y_n)$,*

$$\hat{G}_n(\phi \circ F) \leq 2L \sum_{i=1}^m \hat{G}_n(F_i),$$

where $\hat{G}_n(\phi \circ F)$ are the gaussian averages of $\phi \circ F$ with respect to the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $\hat{G}_n(F_i)$ are the gaussian averages of F_i with respect to the sample $X_1, \ldots, X_n$.

Proof Without loss of generality, we may assume that each class F_i is finite, denote by $|F_i|$ its cardinality and let f_k^i be the k-th element in F_i. Let Γ be a multi-index set $\Gamma = \{(j_1, ..., j_m) : 1 \leq j_i \leq |F_i|\}$. Hence, there is a one-to-one correspondence between F and Γ, which is given by $\alpha = (j_1, \ldots, j_m) \mapsto f_\alpha = (f_{j_1}^1, \ldots, f_{j_m}^m)$. For every $\alpha = (j_1, \ldots, j_m) \in \Gamma$, let

$$X_\alpha = \sum_{k=1}^n \phi\big(y_k, f_\alpha(x_k)\big) g_k,$$

and

$$Y_\alpha = L \sum_{i=1}^m \sum_{k=1}^n f_{j_i}^i(x_k) h_{ik}$$

where (g_k) and (h_{ik}) are all standard independent normal random variables. It is easy to see that for every $\alpha, \alpha' \in \Gamma$,

$$\begin{aligned}
\|X_\alpha - X_{\alpha'}\|_2^2 &= \sum_{k=1}^n \Big(\phi\big(y_k, f_\alpha(x_k)\big) - \phi\big(y_k, f_{\alpha'}(x_k)\big)\Big)^2 \\
&\leq L^2 \sum_{i=1}^m \sum_{k=1}^n \big(f_{j_i}^i(x_k) - f_{j_i'}^i(x_k)\big)^2 \\
&= \|Y_\alpha - Y_{\alpha'}\|_2^2.
\end{aligned}$$

Our claim follows from Slepian's Lemma and the observation that $\mathbf{E}\sup_\alpha X_\alpha = n\hat{G}_n(\phi \circ F)$ and that $\mathbf{E}\sup_\alpha Y_\alpha = nL\sum_{i=1}^m \hat{G}_n(F_i)$. ■

Corollary 15 *Let $\mathcal{A}, F, F_1, \ldots, F_m, \phi$ be as in Theorem 14. Consider a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{A} \to [0,1]$ and suppose that $\phi : \mathcal{Y} \times \mathcal{A} \to [0,1]$ dominates $\mathcal{L}$. Then, for any integer n there is a probability of at least $1-\delta$ that every f in F has*

$$\mathbf{E}\mathcal{L}(Y, f(X)) \leq \hat{\mathbf{E}}_n\phi(Y, f(X)) + cL\sum_{j=1}^m G_n(F_j) + \sqrt{\frac{8\ln(2/\delta)}{n}}\,.$$

3.3 Boolean Combinations of Functions

Theorem 16 *For a fixed boolean function $g : \{\pm 1\}^k \to \{\pm 1\}$ and classes $F_1, \ldots, F_k$ of $\{\pm 1\}$-valued functions,*

$$G_n(g(F_1, \ldots, F_k)) \leq 2\sum_{j=1}^k G_n(F_j).$$

Proof First, we extend the boolean function g to a function $g : \mathbb{R}^k \to [-1,1]$ as follows: for $x \in \mathbb{R}^k$, define $g(x) = (1 - \|x-a\|)g(a)$ if $\|x-a\| < 1$ for some $a \in \{\pm 1\}^k$, and $g(x) = 0$ otherwise. The function is well-defined since all pairs of points in the k-cube are separated by distance at least 2. Clearly, $g(0) = 0$, and g is Lipschitz with constant 1. The theorem follows from Theorem 14, with $m = k$ and $\phi = g$. ■

4. Examples

The error bounds presented in previous sections can be used as the basis of a complexity regularization algorithm for model selection. This algorithm minimizes an upper bound on error involving the sample average of a cost function and a gaussian or Rademacher complexity penalty term. We have seen that these upper bounds in terms of gaussian and Rademacher complexities can never be significantly worse than bounds based, for example, on combinatorial dimensions. They can have a significant advantage over such bounds, since they measure the complexity of the class on the training data, and hence can reflect the properties of the particular probability distribution that generates the data. The computation of these complexity penalties involves an optimization over the model class. The structural results of the previous section give a variety of techniques that can simplify this optimization problem. For example, voting methods involve optimization over the convex hull of some function class H. By Theorem 12 part 2, we can estimate $G_n(\mathrm{conv}H)$ by solving a maximization problem over the base class H. In this section, we give some other examples illustrating this approach. In all cases, the resulting error bounds decrease at least as fast as $1/\sqrt{n}$.

4.1 Decision Trees

A binary-valued decision tree can be represented as a fixed boolean function of the decision functions computed at its nodes. Theorem 16 implies that the gaussian complexity of the class of decision trees of a certain size can be bounded in terms of the gaussian complexity of the class of node decision functions. Typically, this is simpler to compute. The following result gives a refinement of this idea, based on the representation (see, for example, Golea et al., 1998) of a decision tree as a thresholded linear combination of the indicator functions of the leaves.

Theorem 17 *Let P be a probability distribution on $\mathcal{X} \times \{-1, 1\}$, and let H be a set of binary-valued functions defined on $\mathcal{X}$. Let T be the class of decision trees of depth no more than d, with decision functions from H. For a training sample $(X_1, Y_1, \ldots, X_n, Y_n)$ drawn from P^n and a decision tree from T, let $\tilde{P}_n(l)$ denote the proportion of all training examples which reach leaf l and are correctly classified. Then with probability at least $1 - \delta$, every decision tree t from T with L leaves has $\Pr(y \neq t(x))$ no more than*

$$\hat{P}_n(y \neq t(x)) + \sum_l \min(\tilde{P}_n(l), cdG_n(H)) + \sqrt{\frac{c \ln(L/\delta)}{2n}} \, .$$

Notice that the key term in this inequality is $O(dLG_n(H))$. It can be considerably smaller if many leaves have small empirical weight. This is the case, for instance, if $G_n(H) = O(n^{-1/2})$ and many leaves have weight less than $O\left(dn^{-1/2} \ln n\right)$.

Proof For a tree of depth d, the indicator function of a leaf is a conjunction of no more than d decision functions. More specifically, if the decision tree consists of decision nodes chosen from a class H of binary-valued functions, the indicator function of leaf l (which takes value 1 at a point x if x reaches l, and 0 otherwise) is a conjunction of d_l functions from H, where d_l is the depth of leaf l. We can represent the function computed by the tree as the sign of

$$f(x) = \sum_l w_l \sigma_l \bigwedge_{i=1}^{d_l} h_{l,i}(x),$$

where the sum is over all leaves l, $w_l > 0$, $\sum_l w_l = 1$, $\sigma_l \in \{\pm 1\}$ is the label of leaf l, $h_{l,i} \in H$, and the conjunction is understood to map to $\{0, 1\}$. Let F be this class of functions. Choose a family $\{\phi_L : L \in \mathbb{N}\}$ of cost functions such that each ϕ_L dominates the step function $\mathbf{1}(yf(x) \leq 0)$ and has a Lipschitz constant L. For each L, Theorem 7 implies that with probability at least $1 - \delta$,

$$\Pr(yf(x) \leq 0) \leq \hat{\mathbf{E}}_n(\phi_L(yf(x))) + 2LR_n(F) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

By setting $\delta_L = 6\delta/(\pi^2 L)$, applying this result to all positive integer values of L, and summing over L, we see that with probability at least $1 - \delta$, every $f \in F$ and every ϕ_L has

$$\Pr(yf(x) \leq 0) \leq \hat{\mathbf{E}}_n(\phi_L(yf(x))) + 2LR_n(F) + \sqrt{\frac{\ln(\pi^2 L/3\delta)}{2n}}.$$

Define $\phi_L(\alpha)$ to be 1 if $\alpha \leq 0$, $1 - L\alpha$ if $0 < \alpha \leq 1/L$, and 0 otherwise, where L will be computed later. Let $\tilde{P}_n(l)$ denote the proportion of training examples which reach leaf l and are correctly classified ($y = \sigma_l$). Then we have

$$\begin{aligned}
&\hat{\mathbf{E}}_n(\phi_L(yf(x))) + 2LR_n(F) \\
&\quad = \hat{P}_n(yf(x) \leq 0) + \sum_l \tilde{P}_n(l)\phi_L(w_l) + 2LR_n(F) \\
&\quad = \hat{P}_n(yf(x) \leq 0) + \sum_l \tilde{P}_n(l)\max(0, 1 - Lw_l) + 2LR_n(F) \\
&\quad = \hat{P}_n(yf(x) \leq 0) + \sum_l \max(0, (1 - Lw_l)\tilde{P}_n(l)) + 2LR_n(F).
\end{aligned}$$

Now, choose $w_l = 0$ for $\tilde{P}_n(l) \leq 2R_n(F)$, and $w_l = 1/L$ otherwise, where $L = |\{l : \tilde{P}_n(l) > 2R_n(F)\}|$. (Notice that choosing $w_l = 0$ for labelled examples for which $yf(x) > 0$ can only increase the bound.) Then we have

$$\begin{aligned}
\hat{P}_n(\phi(yf(x))) + 2LR_n(F) \quad \leq \quad & \hat{P}_n(yf(x) \leq 0) \\
& + \sum_l \mathbf{1}\left(\tilde{P}_n(l) \leq 2R_n(F)\right)\tilde{P}_n(l) \\
& + 2R_n(F)\sum_l \mathbf{1}\left(\tilde{P}_n(l) > 2R_n(F)\right) \\
= \quad & \hat{P}_n(yf(x) \leq 0) + \sum_l \min(\tilde{P}_n(l), 2R_n(F)).
\end{aligned}$$

Theorem 12 part 2, Theorem 16, and Lemma 4 together imply that

$$R_n(F) \leq C\left(dG_n(H) + \frac{1}{n}\right)\ln n,$$

which implies the result. ■

4.2 Neural Networks

Neural network methods (see, for example, Anthony and Bartlett, 1999) use repeated compositions of linear functions with scalar nonlinearities, $\sigma : \mathbb{R} \to [-1, 1]$, where σ is typically monotonic and smooth. The following theorem bounds the gaussian complexity of a two-layer neural network with constraints on the magnitudes of the weights.

Theorem 18 *Suppose that $\sigma : \mathbb{R} \to [-1, 1]$ has Lipschitz constant L and satisfies $\sigma(0) = 0$. Define the class computed by a two-layer neural network with 1-norm weight constraints as*

$$F = \left\{ x \mapsto \sum_i w_i\sigma\left(v_i \cdot x\right) : \|w\|_1 \leq 1,\ \|v_i\|_1 \leq B \right\}.$$

Then for $x_1, \ldots, x_n$ in $\mathbb{R}^k$,

$$\hat{G}_n(F) \leq \frac{cLB(\ln k)^{1/2}}{n} \max_{j,j'} \sqrt{\sum_{i=1}^{n} \left(x_{ij} - x_{ij'}\right)^2},$$

where $x_i = (x_{i1}, \ldots, x_{ik})$.

It is straightforward to extend this result to networks with more than two layers, and to networks with multiple outputs. The theorem is immediate from the following result for bounded linear functions.

Lemma 19 *For $x \in \mathbb{R}^k$, define*

$$F_1 = \left\{ x \mapsto w \cdot x : w \in \mathbb{R}^k, \|w\|_1 \leq 1 \right\}.$$

For any $x_1, \ldots, x_n \in \mathbb{R}^k$ we have

$$\hat{G}_n(F_1) \leq \frac{c}{n} (\ln k)^{1/2} \max_{j,j'} \left(\sum_{i=1}^{n} (x_{ij} - x_{ij'})^2 \right)^{1/2}.$$

The proof uses the following inequality for gaussian processes which follows from Slepian's Lemma (see, for example, Ledoux and Talagrand, 1991, Pisier, 1989).

Lemma 20 *Let $Z_1, \ldots, Z_k$ be random variables such that for every $1 \leq j \leq k$, $Z_j = \sum_{i=1}^{n} a_{ij} g_i$, where $g_1, \ldots, g_n$ are independent gaussian $N(0,1)$ random variables. Then there is an absolute constant c such that*

$$\mathbf{E} \max_{1 \leq j \leq k} Z_j \leq c(\ln k)^{1/2} \max_{j,j'} \sqrt{\mathbf{E}(Z_j - Z_{j'})^2}.$$

Proof *(of Lemma 19)* From the definitions, $\hat{G}_n(F_1)$ is equal to

$$\begin{aligned} \mathbf{E} \sup_{f \in F} \frac{2}{n} \sum_{i=1}^{n} g_i f(x_i) &= \mathbf{E} \sup_{w : \|w\|_1 \leq 1} \frac{2}{n} \sum_{i=1}^{n} g_i w \cdot x_i \\ &= \mathbf{E} \sup_{w : \|w\|_1 \leq 1} w \cdot \frac{2}{n} \sum_{i=1}^{n} g_i x_i \,. \end{aligned}$$

Clearly, this inner product is maximized when w is at one of the extreme points of the ℓ_1 ball, which implies

$$\hat{G}_n(F_1) = \mathbf{E} \max_j \frac{2}{n} \sum_{i=1}^{n} g_i x_{ij},$$

where $x_i = (x_{i1}, \ldots, x_{ik})$. Note that we can write

$$\hat{G}_n(F_1) = \frac{2}{n} \mathbf{E} \max_j Z_j$$

where $Z_j = \sum_{i=1}^n g_i x_{ij}$. Since each Z_j is gaussian, we can apply Slepian's Lemma to obtain

$$\begin{aligned}\hat{G}_n(F_1) &\leq \frac{2c}{n}(\ln k)^{1/2} \max_{j,j'} \sqrt{\mathbf{E}(Z_j - Z_{j'})^2} \\ &= \frac{2c}{n}(\ln k)^{1/2} \max_{j,j'} \sqrt{\mathbf{E}\left(\sum_{i=1}^n g_i(x_{ij} - x_{ij'})\right)^2} \\ &= \frac{2c}{n}(\ln k)^{1/2} \max_{j,j'} \sqrt{\sum_{i=1}^n (x_{ij} - x_{ij'})^2}.\end{aligned}$$

■

4.3 Kernel Methods

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on a compact space $\mathcal{X}$ is a continuous function such that for all $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathcal{X}$, the Gram matrix K, with $K_{ij} = k(x_i, x_j)$, is positive semi-definite and symmetric. Kernel methods, such as support vector machines (see, for example Cristianini and Shawe-Taylor, 2000) use kernel expansions of the form

$$x \mapsto \sum_{i=1}^n \alpha_i k(x_i, x).$$

These methods typically restrict $\alpha = (\alpha_1, \ldots, \alpha_n)$ so that $\alpha' K \alpha$ is small. The following theorem gives a margin-based estimate of misclassification probability for these functions.

Theorem 21 *Fix $B, \gamma > 0$, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with*

$$\sup_{x \in \mathcal{X}} |k(x, x)| < \infty.$$

Define the margin cost function $\phi : \mathbb{R} \to [0, 1]$ as

$$\phi(\alpha) = \begin{cases} 1 & \text{if } \alpha \leq 0 \\ 1 - \alpha/\gamma & \text{if } 0 < \alpha \leq \gamma \\ 0 & \text{if } \alpha > \gamma. \end{cases}$$

Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are chosen independently according to some probability distribution P on $\mathcal{X} \times \{\pm 1\}$. Then with probability at least $1 - \delta$, every function f of the form

$$f(x) = \sum_{i=1}^n \alpha_i k(X_i, x)$$

with $\sum_{i,j} \alpha_i \alpha_j k(X_i, X_j) \leq B^2$ satisfies

$$P(Yf(X) \leq 0) \leq \hat{\mathbf{E}}_n \phi(Yf(X)) + \frac{4B}{\gamma n} \sqrt{\sum_{i=1}^n k(X_i, X_i)} + \left(\frac{8}{\gamma} + 1\right) \sqrt{\frac{\ln(4/\delta)}{2n}}.$$

To prove this theorem, we need to recall some properties of kernel expansions. To every kernel k we can associate a feature map $\Phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and for all $x_1, x_2 \in \mathcal{X}$, $k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$. If $\| \cdot \|$ denotes the norm in $\mathcal{H}$, we have

$$\left\| \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right\|^2 = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j),$$

and hence

$$\begin{aligned} F &= \left\{ x \mapsto \sum_{i=1}^{m} \alpha_i k(x, x_i) : m \in \mathbb{N},\ x_i \in \mathcal{X},\ \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \leq B^2 \right\} \\ &\subseteq \left\{ x \mapsto \langle w, \Phi(x) \rangle : \|w\| \leq B \right\}. \end{aligned}$$

The following lemma, combined with Theorem 7 and Theorem 11, implies the theorem.

Lemma 22 *Suppose that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel, and let $X_1, \ldots, X_n$ be random elements of $\mathcal{X}$. Then for the class F defined above,*

$$\hat{G}_n(F) \leq \frac{2B}{n} \sqrt{\sum_{i=1}^{n} k(X_i, X_i)},$$

$$\hat{R}_n(F) \leq \frac{2B}{n} \sqrt{\sum_{i=1}^{n} k(X_i, X_i)}.$$

Proof Suppose that $\mathcal{H}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$, and the kernel k has feature map $\Phi : \mathcal{X} \to \mathcal{H}$. Let $g_1, \ldots, g_n$ be independent standard normal random variables. Then

$$\begin{aligned} \hat{G}_n(F) &\leq \mathbf{E} \left[\sup_{\|w\| \leq B} \left\langle w, \frac{2}{n} \sum_{i=1}^{n} g_i \Phi(X_i) \right\rangle \middle| X_i \right] \\ &= \frac{2B}{n} \mathbf{E} \left[\left\| \sum_{i=1}^{n} g_i \Phi(X_i) \right\| \middle| X_i \right] \\ &= \frac{2B}{n} \mathbf{E} \left[\left(\sum_{i,j} g_i g_j k(X_i, X_j) \right)^{1/2} \middle| X_i \right] \\ &\leq \frac{2B}{n} \left(\sum_{i,j} \mathbf{E} \left[g_i g_j k(X_i, X_j) \middle| X_i \right] \right)^{1/2} \\ &= \frac{2B}{n} \left(\sum_{i} \mathbf{E} \left[g_i^2 k(X_i, X_i) \middle| X_i \right] \right)^{1/2} \\ &= \frac{2B}{n} \left(\sum_{i} k(X_i, X_i) \right)^{1/2}, \end{aligned}$$

where the second inequality is Jensen's (and it is easy to see that the first inequality is an equality).

Clearly, the same argument applies with any independent, zero mean, unit variance random variables replacing the g_i, which gives the same bound for $\hat{R}_n(F)$. ■

From the definitions and Jensen's inequality,

$$R_n(F) = \mathbf{E}\hat{R}_n(F) \leq 2B\sqrt{\frac{\mathbf{E}k(X,X)}{n}}$$
$$G_n(F) = \mathbf{E}\hat{G}_n(F) \leq 2B\sqrt{\frac{\mathbf{E}k(X,X)}{n}}$$

Notice that $\mathbf{E}k(X,X)$ is the trace (sum of the eigenvalues) of the integral operator T_k on $L_2(\mu)$,

$$T_k(f) = \int k(x,y)f(y)d\mu(y),$$

where μ is the induced probability measure on $\mathcal{X}$.

Acknowledgments

Thanks to Arthur Gretton, Jonathan Baxter and Gábor Lugosi for helpful discussions, to the anonymous reviewers for suggestions, and especially to the reviewer who suggested a substantially simpler proof of Theorem 14. This work was partially supported by the Australian Research Council.

Appendix A. Proof of Lemma 3

The expected maximum discrepancy, $D_n(F)$, measures the average difference between function values on two fixed subsets of the data. The Rademacher complexity, $R_n(F)$, measures the difference on two randomly chosen subsets. The idea behind the proof of the first part of the lemma is to show that the size of the random subsets is very close to their expectation, and, because the data is independent, all choices of these equally sized subsets are equivalent.

Define

$$s(N) = \frac{2}{n}\mathbf{E}\left[\sup_{f\in F}\sum_{i=1}^{n}\sigma_i f(X_i)\middle|\sum_{i=1}^{n}\sigma_i = N\right].$$

Then we have

$$\begin{aligned} R_n(F) &= \mathbf{E} \sup_{f \in F} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right| \\ &\geq \mathbf{E} \sup_{f \in F} \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \\ &= \mathbf{E}\mathbf{E} \left[\sup_{f \in F} \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \middle| \sum_{i=1}^{n} \sigma_i \right] \\ &= \mathbf{E} s \left(\sum_{i=1}^{n} \sigma_i \right), \end{aligned}$$

where the inequality is an equality when $f \in F$ implies $-f \in F$. It is easy to see (from the independence of the X_i) that

$$D_n(F) = s(0) = s\left(\mathbf{E} \sum_{i=1}^{n} \sigma_i \right).$$

Furthermore, s satisfies a Lipschitz condition. To see this, choose n_1, n_2 satisfying $0 \leq n_2 < n_1 \leq n$ and write

$$\begin{aligned} s(n_1) &= \frac{2}{n} \mathbf{E} \sup \left(\sum_{i=1}^{n/2-n_1/2} (f(X_{2i}) - f(X_{2i-1})) + \sum_{i=n-n_1+1}^{n} f(X_i) \right), \\ s(n_2) &= \frac{2}{n} \mathbf{E} \sup \left(\sum_{i=1}^{n/2-n_1/2} (f(X_{2i}) - f(X_{2i-1})) + \sum_{i=n-n_1+1}^{n} f(X_i) \right. \\ &\qquad \left. + \sum_{i=n/2-n_1/2+1}^{n/2-n_2/2} (f(X_{2i}) - f(X_{2i-1})) - \sum_{i=n-n_1+1}^{n-n_2} f(X_i) \right). \end{aligned}$$

Clearly, the two expressions differ only in the last two terms inside the supremum in the expression for $s(n_2)$. Each of these has magnitude no more than $(n_2 - n_1)$. Thus,

$$|s(n_1) - s(n_2)| \leq \frac{4|n_2 - n_1|}{n}.$$

Thus, if $N = \sum_i \sigma_i$,

$$\begin{aligned} \Pr\left(|s(N) - s(\mathbf{E}N)| \geq \epsilon \right) &\leq \Pr\left(|N - \mathbf{E}N| > \frac{\epsilon n}{4} \right) \\ &\leq 2 \exp\left(-\frac{\epsilon^2 n}{32} \right), \end{aligned}$$

by Chernoff's inequality. A standard integration (see, for example, Devroye et al., 1996, p208) shows that

$$|\mathbf{E}s(N) - s(\mathbf{E}N)| \leq \mathbf{E}|s(N) - s(\mathbf{E}N)| \leq 4\sqrt{\frac{2}{n}}.$$

Since $R_n(F) \geq \mathbf{E}s(N)$, this proves the upper bound on $D_n(F)$.

Define $-F = \{-f : f \in F\}$. If $F = -F$, $R_n(F) = \mathbf{E}s(N)$, so

$$R_n(F) = R_n(F \cup -F) \leq D_n(F \cup -F) + 4\sqrt{\frac{2}{n}},$$

which is the required lower bound if F is closed under negation. In general, it is easy to see that $D_n(F \cup -F) \leq 2D_n(F)$.

The final part of the lemma follows immediately from McDiarmid's inequality.

Appendix B. Proof of Theorem 5

We set $\mathcal{L}(Y, f(X)) = \mathbf{1}(Y \neq f(X))$ and proceed as in the proof of Theorem 8. For all $f \in F$,

$$P(Y \neq f(X)) = \mathbf{E}\mathcal{L}(Y, f(X)) \leq \hat{\mathbf{E}}_n\mathcal{L}(Y, f(X)) + \sup_{h \in \mathcal{L} \circ F} \left(\mathbf{E}h - \hat{\mathbf{E}}_n h\right)$$

In this case, when (X_i, Y_i) changes, the supremum changes by no more than $1/n$, so McDiarmid's inequality implies that with probability at least $1 - \delta$, every $f \in F$ satisfies

$$\mathbf{E}\mathcal{L}(Y, f(X)) \leq \hat{\mathbf{E}}_n\mathcal{L}(Y, f(X)) + \mathbf{E}\sup_{h \in \mathcal{L} \circ F} \left(\mathbf{E}h - \hat{\mathbf{E}}_n h\right) + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

The same argument as in the proof of Theorem 8 shows that

$$\begin{aligned}
\mathbf{E}\sup_{h \in \mathcal{L} \circ F} \left(\mathbf{E}h - \hat{\mathbf{E}}_n h\right) &\leq \mathbf{E}\sup_{h \in \mathcal{L} \circ F} \frac{2}{n}\sum_{i=1}^{n} \sigma_i h(X_i, Y_i) \\
&= \mathbf{E}\sup_{f \in F} \frac{2}{n}\sum_{i=1}^{n} \sigma_i \mathbf{1}(Y_i \neq f(X_i)) \\
&= \mathbf{E}\sup_{f \in F} \frac{2}{n}\sum_{i=1}^{n} \sigma_i (1 - Y_i f(X_i))/2 \\
&= \mathbf{E}\sup_{f \in F} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i) \\
&= \frac{R_n(F)}{2},
\end{aligned}$$

where we have used the fact that $Y_i, f(X_i) \in \{\pm 1\}$, and that the conditional distribution of $\sigma_i Y_i$, given Y_i, is the same as the distribution of σ_i.

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.

Peter 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(2):525–536, 1998.

Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, pages 224–240, 2001.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Methods*. Cambridge University Press, 2000.

Luc Devroye, Lázsló Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Applications of Mathematics: Stochastic Modelling and Applied Probability (31). Springer, 1996.

Mostefa Golea, Peter L. Bartlett, and Wee Sun Lee. Generalization in decision trees and DNF: Does size matter? In *NIPS 10*, pages 259–265, 1998.

Michael J. Kearns, Yishay Mansour, Andrew Y. Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.

V. Koltchinskii. Rademacher penalties and structural risk minimization. Technical report, Department of Mathematics and Statistics, University of New Mexico, 2000.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. Technical report, Department of Mathematics and Statistics, University of New Mexico, 2000a.

V. I. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In E. Gine, D. Mason, and J. Wellner, editors, *High Dimensional Probability II*, pages 443–459. 2000b.

E. B. Kong and T. G. Dietterich. Error-correcting output coding corrects bias and variance. In *Proc. 12th International Conference on Machine Learning*, pages 313–321. Morgan Kaufmann, 1995.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

Llew Mason, Peter L. Bartlett, and Jonathan Baxter. Improved generalization through explicit optimization of margins. *Machine Learning*, 38(3):243–255, 2000.

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

Shahar Mendelson. l-norm and its application to learning theory. *Positivity*, 2001a. (To appear—see http://www.axiom.anu.edu.au/∼shahar.).

Shahar Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. (To appear, IEEE Transactions on Information Theory; see http://www.axiom.anu.edu.au/∼shahar), 2001b.

G. Pisier. *The volume of convex bodies and Banach space geometry*. Cambridge University Press, 1989.

Robert E. Schapire. Using output codes to boost multiclass learning problems. In *Machine Learning: Proc. Fourteenth International Conference*, pages 313–321, 1997.

Robert E. Schapire, Yoav Freund, Peter L. Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26(5): 1651–1686, October 1998.

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimisation over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

N. Tomczak-Jaegermann. *Banach-Mazur distance and finite-dimensional operator ideals*. Number 38 in Pitman Monographs and Surveys in Pure and Applied Mathematics. Pitman, 1989.

Vladimir N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16 (2):264–280, 1971.