Journal of Machine Learning Research 2 (2002) 527-550 Submitted 6/2001; Published 3/2002

Covering Number Bounds of Certain Regularized Linear Function Classes

Tong Zhang TZHANG@WATSON.IBM.COM
T.J. Watson Research Center
Route 134, Yorktown Heights, NY 10598, U.S.A.

Editor: Peter L. Bartlett

Abstract

Recently, sample complexity bounds have been derived for problems involving linear functions such as neural networks and support vector machines. In many of these theoretical studies, the concept of covering numbers played an important role. It is thus useful to study covering numbers for linear function classes. In this paper, we investigate two closely related methods to derive upper bounds on these covering numbers. The first method, already employed in some earlier studies, relies on the so-called Maurey's lemma; the second method uses techniques from the mistake bound framework in online learning. We compare results from these two methods, as well as their consequences in some learning formulations.

Keywords: Covering Numbers, Learning Sample Complexity, Sparse Approximation, Mistake Bounds

1. Introduction

Assume we have a set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with corresponding desired output variables $y_1, \ldots, y_n$. The task of supervised learning is to estimate the functional relationship $y \approx q(\mathbf{x})$ between the input variable $\mathbf{x}$ and the output variable y from the training examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.

A simple and useful model of an input-output functional relationship is to assume that the output variable can be approximately expressed as a linear combination of its input vector components. With appropriate (nonlinear) features, linear models can be used to approximate an arbitrary nonlinear function. One useful technique for constructing nonlinear features is *kernel methods*, where each feature is a function of the current input and one of the example inputs (such as their distance). If a kernel function is positive definite, then the sample space feature representation is also equivalent to an implicit representation in the kernel associated reproducing kernel Hilbert space, which can be infinite dimensional. Kernel methods have been successfully employed in methods such as support vector machines and Gaussian processes (Cristianini and Shawe-Taylor, 2000). Furthermore, one can linearize an arbitrary nonlinear model such as a neural network by using weighted averaging over all possible neural networks in the model. This approach of model combination (also called committee) has been widely used in machine learning to improve the predictive

performance of a nonlinear model. For example, the recently proposed boosting algorithm (see Freund and Schapire, 1997) can be considered as an implementation of this idea.

It is therefore useful to study the generalization performance of linear prediction methods. From the computational learning theory point of view, such performance measurements, or sample complexity bounds, can be described by a quantity called *covering number* (see Pollard, 1984, Vapnik, 1998), which measures the size of a parametric function family. For a two-class classification problem, its covering number can be bounded by using a combinatorial quantity called *VC-dimension* (Sauer, 1972, Vapnik and Chervonenkis, 1971). More recently, researchers have discovered other combinatorial quantities (or *dimensions*) that are useful for bounding covering numbers. Consequently, the concept of VC-dimension has been generalized to more general problems (Devroye et al., 1996, Pollard, 1984, Vapnik, 1998).

In a linear prediction model, we assume that the input-output functional relationship can be expressed as $y \approx \mathbf{w} \cdot \mathbf{x}$, where $\mathbf{w} \cdot \mathbf{x}$ denotes the inner product of vectors $\mathbf{w}$ and $\mathbf{x}$. The prediction quality of this model can be measured by a loss function $\mathcal{L}$, and our goal is to find a linear weight $\mathbf{w}$ from the training data so that it minimizes the expected loss:

$$\begin{aligned} &\min E_{\mathbf{X},y}\mathcal{L}(\mathbf{w} \cdot \mathbf{x}, y) \\ &\text{s.t.} \qquad g(\mathbf{w}) \leq A. \end{aligned} \tag{1}$$

$E_{\mathbf{X},y}$ denotes the expectation over an unknown distribution D on $(\mathbf{x}, y)$. In supervised learning, we often assume that the training data are independent samples from D. The constraint $g(\mathbf{w}) \leq A$ limit the size of the underlying linear hypothesis family. This condition balances the prediction power and the learning complexity of the family, and is widely used in many recent linear prediction methods such as Gaussian process, support vector machines and boosting. By introducing an appropriately chosen Lagrangian multiplier $\lambda \geq 0$ for the constraint $g(\mathbf{w}) \leq A$, the minimization of (1) is equivalent to minimizing

$$E_{\mathbf{X},y}\mathcal{L}(\mathbf{w} \cdot \mathbf{x}, y) + \lambda g(\mathbf{w}).$$

This equivalent formulation occurs more frequently in practical applications. Usually $E_{\mathbf{X},y}$ is replaced by the empirical expectation over the training data, and the regularization parameter λ is determined by cross-validation. Similarly, if $\lambda > 0$, then we can regard $1/\lambda$ as a Lagrangian multiplier for the constraint $E_{\mathbf{X},y}\mathcal{L}(\mathbf{w} \cdot \mathbf{x}, y) \leq s$, which leads to the third equivalent formulation as follows:

$$\begin{aligned} &\min g(\mathbf{w}) \\ &\text{s.t.} \qquad E_{\mathbf{X},y}\mathcal{L}(\mathbf{w} \cdot \mathbf{x}, y) \leq s. \end{aligned}$$

For a regression problem, we often choose a squared loss function $\mathcal{L}(q, y) = (y - q)^2$. For a binary classification problem where $y = \pm 1$, the linear decision rule with $\mathbf{w}$ is:

$$c(\mathbf{w}, \mathbf{x}) = \begin{cases} 1 & \text{if} \quad \mathbf{w} \cdot \mathbf{x} > 0, \\ -1 & \text{if} \quad \mathbf{w} \cdot \mathbf{x} \leq 0. \end{cases}$$

The loss function is the classification error $\mathcal{L}(\mathbf{w} \cdot \mathbf{x}, y) = |y - c(\mathbf{w}, \mathbf{x})|/2$.

Note that in the literature, one often encounters a more general type of linear functional: $\mathbf{w} \cdot \mathbf{x} + b$, where b is called *bias*. However, one can easily convert this formulation into one in which b is zero. This is achieved by letting $\tilde{\mathbf{x}} = [\mathbf{x}, 1]$, and $\tilde{\mathbf{w}} = [\mathbf{w}, b]$: $\tilde{\mathbf{w}} \cdot \tilde{\mathbf{x}} = \mathbf{w} \cdot \mathbf{x} + b$. Therefore we assume a linear form with $b = 0$ throughout this paper.[1]

Since the classification error function is non-convex which may cause computational problems, one often employs a convex upper bound of classification error as loss function in the training. For example, we may consider $\mathcal{L}(q, y) = \log_2(1 + \exp(-qy))$ which leads to logistic regression.

Since the complexity regularization condition $g(\mathbf{w}) \leq A$ in (1) determines the shape of the underlying linear function classes, it has a significant impact on the generalization ability of (1) or its equivalent formulations. As an example, a support vector machine in its primal formulation can be regarded as a special case of (1) with square regularization condition $g(\mathbf{w}) = \mathbf{w} \cdot \mathbf{w}$. For data with bounded 2-norms, Vapnik has shown that the VC dimension of a linear function class bounded in 2-norm (assume it separates the input data with a positive margin) is independent of the input space dimension. He then argued that the generalization performance of a support vector machine does not depend on the input data dimension. This observation is significant since it means that with an appropriate regularization on the parameter space, the input data dimension does not have an adverse impact on the ability to learn from data. This prevents the so called *curse-of-dimension* in many learning formulations. One natural question to ask is whether this property is unique to the 2-norm regularization. For example, what is the implication of using some other regularization conditions in (1)? The answer to this question can be of great interest since people have already used different kinds of non 2-norm regularization terms in engineering applications.

Related to this question, there have been a number of recent works on large margin linear classification using non 2-norm regularization. For example, Bartlett (1998) studied the performance of neural networks under the 1-norm regularization of the associated weights. The same idea has also been applied by Schapire et al. (1998) to analyze the boosting algorithm. It has later been realized that these theoretical results are directly related to some newly obtained covering number bounds for linear function classes under appropriate regularization conditions. Consequently, a number of studies have appeared in the last few years on covering numbers for linear function classes (Anthony and Bartlett, 1999, Guo et al., 1999, Gurvits, 1997, Williamson et al., 1999, 2000).

In Section 3, we derive new covering number bounds, which complement and improve results from previous studies. In our analysis, we emphasize the importance of covering number bounds that do not depend on the input data dimension. Based on these new covering number results, generalization performance of formulation (1) is obtained in the PAC learning framework. Specifically, under certain non 2-norm regularization conditions, weight $\mathbf{w}$ computed with (1) can also lead to generalization performances that do not deteriorate when the input dimension increases. Note that this property has been regarded as a major theoretical advantage for support vector machines.

The paper is organized as follows. In Section 2, we briefly review the concept of covering numbers as well as some main results for analyzing the performance of a learning

1. The transformation may lead to a slightly different optimization problem in that b is regularized after the transformation but may not be so beforehand. However, the practical difference is not significant.

algorithm. In Section 3, we introduce the regularization idea. Our main goal is to construct regularization conditions for linear function classes so that the resulting covering numbers are independent of the input data dimension. We also introduce a new technique of using online learning to prove covering number bounds. We show that this method leads to results that improve some previous bounds. Section 4 applies the covering number bounds to analyze some specific examples of (1). Section 5 summarizes results obtained in this paper.

2. Covering Numbers

We formulate the statistical learning problem as to find a parameter from random observations to minimize the expected loss (*risk*): given a loss function $\mathcal{L}(\alpha, \mathbf{x})$ and n observations $X_1^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ independently drawn from a fixed but unknown distribution D, we want to find α that minimizes the true risk defined as:

$$R(\alpha) = E_{\mathbf{x}} \mathcal{L}(\alpha, \mathbf{x}) = \int \mathcal{L}(\alpha, \mathbf{x}) \, dP_D(\mathbf{x}), \tag{2}$$

where $E_{\mathbf{x}}$ denotes the expectation over the unknown distribution D. In order to make the discussion more general, we have adopted different notations than those in (1). In particular, y in (1) is absorbed into the input variable $\mathbf{x}$ in (2); the linear weight parameter $\mathbf{w}$ in (1) corresponds to the general parameter α in (2).

Without any assumption of the underlying distribution D on $\mathbf{x}$, a natural method for solving (2) with a limited number of observations is the *empirical risk minimization* (ERM) method (Vapnik, 1998). We choose a parameter α that minimizes the observed risk:

$$R(\alpha, X_1^n) = E_{X_1^n} \mathcal{L}(\alpha, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\alpha, \mathbf{x}_i),$$

where we use $E_{X_1^n}$ to denote the empirical expectation over the observed data.

The learning behavior of this method with a finite sample size can be studied under the VC theory, which relies on the uniform convergence of the empirical risk to the true risk (also called the uniform law of large numbers). Such a uniform convergence bound can be obtained from quantities that measure the size of a *Glivenko-Cantelli* class. For a function class containing a finite number of indices, its size is simply measured by its cardinality. For a general function class, a well known quantity to measure its size, which determines the degree of uniform convergence, is the *covering number*. The covering number concept can be dated back to Pontriagin and Schnirelmann (1932), Kolmogorov (1956), Kolmogorov and Tihomirov (1961): one discretizes (the discretization process can depend on the observation X_1^n) the parameter space into N values $\alpha_1, \ldots, \alpha_N$ so that each $\mathcal{L}(\alpha, \cdot)$ can be approximated by $\mathcal{L}(\alpha_i, \cdot)$ for some i. We shall only describe a simplified version relevant to our purpose.

Definition 1 *Let B be a metric space with metric ρ. Given observations $X_1^n = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, and vectors $f(\alpha, X_1^n) = [f(\alpha, \mathbf{x}_1), \ldots, f(\alpha, \mathbf{x}_n)] \in B^n$ parameterized by α, the covering number in p-norm, denoted as $\mathcal{N}_p(f, \epsilon, X_1^n)$, is the minimum number m of a collection of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m \in B^n$, such that $\forall \alpha$, $\exists \mathbf{v}_j$:*

$$\|\rho(f(\alpha, X_1^n), \mathbf{v}_j)\|_p = \left[\sum_{i=1}^{n} \rho(f(\alpha, \mathbf{x}_i), \mathbf{v}_j^i)^p \right]^{1/p} \leq n^{1/p} \epsilon,$$

where $\mathbf{v}_j^i$ *is the* i*-th component of vector* $\mathbf{v}_j$*. We also define* $\mathcal{N}_p(f, \epsilon, n) = \sup_{X_1^n} \mathcal{N}_p(f, \epsilon, X_1^n)$.

Note that from the definition and Jensen's inequality, we have $\mathcal{N}_p \leq \mathcal{N}_q$ for $p \leq q$. We implicitly assume that the metric on the real line $\mathcal{R}$ is $|x_1 - x_2|$ unless otherwise specified.

The following theorem, which bounds the rate of uniform convergence of a function class in terms of its covering number, is due to Pollard (1984):

Theorem 1 (Pollard 1984) $\forall \epsilon > 0$*, and distribution* D*,*

$$P\left[\sup_{\alpha} |R(\alpha, X_1^n) - R(\alpha)| > \epsilon\right] \leq 8E[\mathcal{N}_1(\mathcal{L}, \epsilon/8, X_1^n)] \exp\left(\frac{-n\epsilon^2}{128M^2}\right),$$

where $M = \sup_{\alpha,\mathbf{x}} \mathcal{L}(\alpha, \mathbf{x}) - \inf_{\alpha,\mathbf{x}} \mathcal{L}(\alpha, \mathbf{x})$*, and* $X_1^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ *are independently drawn from* D*.*

Constants in the above theorem can be improved for certain problems (Dudley, 1984, Haussler, 1989, Vapnik, 1998). However, they yield similar bounds. A result that is more relevant to our purpose is a lemma by Bartlett (1998), where the 1-norm covering number of the function class in Theorem 1 is replaced by an ∞-norm covering number. The latter quantity can be bounded by a scale-sensitive combinatorial dimension (Alon et al., 1997, Gurvits, 1997). Under certain circumstances, these results can replace Theorem 1 to give better estimates.

However, Bartlett's lemma is only for binary-valued function classes. We will thus extend the result into a form that becomes comparable to Theorem 1. In the following theorem, we replace the "margin" concept for classification problems by a notion of separation for general problems. We also avoid introducing the concept of "fat-shattering" dimension which leads to some complicated technical manipulations (Bartlett, 1998). There are two major differences between the following theorem and Theorem 1: 1. with the existence of a γ-separating function, we are able to use different accuracies γ and ϵ respectively in the covering number estimate and in the Chernoff bound; 2. the covering number used in Theorem 2 does not directly correspond to that of the overall loss function.

Theorem 2 *Let* f_1 *and* f_2 *be two functions:* $\mathcal{R}^n \to [0, 1]$ *such that* $|y_1 - y_2| \leq \gamma$ *implies* $f_1(y_1) \leq f_3(y_2) \leq f_2(y_1)$ *where* $f_3 : \mathcal{R}^n \to [0, 1]$ *is a reference separating function, then*

$$P\left[\sup_{\alpha}[E_{\mathbf{x}} f_1(\mathcal{L}(\alpha, \mathbf{x})) - E_{X_1^n} f_2(\mathcal{L}(\alpha, \mathbf{x}))] > \epsilon\right] \leq 4E[\mathcal{N}_\infty(\mathcal{L}, \gamma, X_1^n)] \exp\left(\frac{-n\epsilon^2}{32}\right).$$

Proof See Appendix A. ■

We say that functions $f_1 : \mathcal{R} \to \mathcal{R}$ and $f_2 : \mathcal{R} \to \mathcal{R}$ have a γ separator if there exists a function $f_3 : \mathcal{R} \to \mathcal{R}$, such that $|y_1 - y_2| \leq \gamma$ implies $f_1(y_1) \leq f_3(y_2) \leq f_2(y_1)$.

Given an arbitrary function f_1 and $\gamma > 0$, one can easily construct f_2 and f_3 such that f_1 and f_2 have a γ-separator f_3. To see this, observe that for a function $f(y) : \mathcal{R}^n \to [0, 1]$, if we define $f^\gamma(y) = \sup_{|z-y|<2\gamma} f(z)$, then $f_1(y) = f(y)$ and $f_2(y) = f^\gamma(y)$ have a γ separator $f_3(y) = f^{\gamma/2}(y)$. Therefore Theorem 2 upper bounds the true expected error of a function $f : \mathcal{R}^n \to [0, 1]$ in terms of the empirical expected error of f^γ. For classification problems,

one usually chooses a formulation with a function f that is non-increasing. In this case we have $f^\gamma(y) = \sup_{|z-y|<2\gamma} f(z) = f(y-2\gamma)$. If f is the step function such that $f(z) = 1$ when $z \leq 0$ and $f(z) = 0$ otherwise (corresponding to the classification error function), then f^γ corresponds to the classification error with a positive margin 2γ. In this special case, Theorem 2 yields the lemma of Bartlett (1998).

Theorem 2 leads to the following PAC style generalization error bound for γ-separable functions f_1 and f_2: $\forall \eta > 0$, with probability of at least $1-\eta$ over the observed data X_1^n, for all α:

$$E_{\mathbf{x}} f_1(\mathcal{L}(\alpha, \mathbf{x})) \leq E_{X_1^n} f_2(\mathcal{L}(\alpha, \mathbf{x})) + \sqrt{\frac{32}{n}\left(\ln 4\mathcal{N}_\infty(\mathcal{L}, \gamma, n) + \ln\frac{1}{\eta}\right)}. \tag{3}$$

If we consider a sequence of functions f_2^γ parameterized by γ, so that f_1 and f_2^γ have a γ separator, then the above PAC bound is valid with f_2 replaced by f_2^γ under the assumption that γ is fixed a priori (data independent). However, using an idea described by Shawe-Taylor et al. (1998), it is not difficult to give a bound that is uniformly valid for all γ, even if γ is chosen according to the observed data:

Corollary 1 *Let f_1 be a function $\mathcal{R} \to \mathcal{R}$. Consider a family of functions $f_2^\gamma : \mathcal{R} \to \mathcal{R}$, parameterized by γ, such that $0 \leq f_1 \leq f_2^\gamma \leq 1$. Assume that for all γ, f_1 and f_2^γ has a γ separator. Assume also that $f_2^\gamma(y) \geq f_2^{\gamma'}(y)$ when $\gamma \geq \gamma'$. Let $\gamma_1 > \gamma_2 > \cdots$ be a decreasing sequence of parameters, and p_i be a sequence of positive numbers such that $\sum_{i=1}^\infty p_i = 1$, then for all $\eta > 0$, with probability of at least $1-\eta$ over data:*

$$E_{\mathbf{x}} f_1(\mathcal{L}(\alpha, \mathbf{x})) \leq E_{X_1^n} f_2^\gamma(\mathcal{L}(\alpha, \mathbf{x})) + \sqrt{\frac{32}{n}\left(\ln 4\mathcal{N}_\infty(\mathcal{L}, \gamma_i, X_1^n) + \ln\frac{1}{p_i\eta}\right)}$$

for all α and γ, where for each fixed γ, we use i to denote the smallest index such that $\gamma_i \leq \gamma$.

Proof The result follows from Theorem 2 and basic probability arguments presented by Shawe-Taylor et al. (1998). $\forall i > 0$, with probability at most $p_i\eta$ over X_1^n, we have

$$E_{\mathbf{x}} f_1(\mathcal{L}(\alpha, \mathbf{x})) > E_{X_1^n} f_2^{\gamma_i}(\mathcal{L}(\alpha, \mathbf{x})) + \sqrt{\frac{32}{n}\left(\ln 4\mathcal{N}_\infty(\mathcal{L}, \gamma_i, X_1^n) + \ln\frac{1}{p_i\eta}\right)}.$$

Summing up over i, with probability at most η over X_1^n,

$$E_{\mathbf{x}} f_1(\mathcal{L}(\alpha, \mathbf{x})) > E_{X_1^n} f_2^{\gamma_i}(\mathcal{L}(\alpha, \mathbf{x})) + \sqrt{\frac{32}{n}\left(\ln 4\mathcal{N}_\infty(\mathcal{L}, \gamma_i, X_1^n) + \ln\frac{1}{p_i\eta}\right)}.$$

for at least one i, which implies the corollary. ■

We can further extend the above corollary which is useful for analyzing the generalization performance of (1).

Corollary 2 *Under the assumptions of Corollary 1. We further assume that there is a sequence of functions $\mathcal{L}_1(\alpha, \mathbf{x}), \mathcal{L}_2(\alpha, \mathbf{x}), \ldots$ with α respectively defined on the parametric spaces $\Gamma_1, \Gamma_2, \ldots$. Let q_i be a sequence of positive numbers such that $\sum_{i=1}^{\infty} q_i = 1$, then for all $\eta > 0$, with probability of at least $1 - \eta$ over data:*

$$E_{\mathbf{X}} f_1(\mathcal{L}_j(\alpha, \mathbf{x})) \leq E_{X_1^n} f_2^{\gamma}(\mathcal{L}_j(\alpha, \mathbf{x})) + \sqrt{\frac{32}{n}\left(\ln 4\mathcal{N}_{\infty}(\mathcal{L}_j, \gamma_i, X_1^n) + \ln \frac{1}{p_i q_j \eta}\right)}$$

for all j, $\alpha \in \Gamma_j$, and $\gamma \in [0, 1]$, where for each fixed γ, we use i to denote the smallest index such that $\gamma_i \leq \gamma$.

Proof Similar to the proof of Corollary 1. $\forall j > 0$, with probability at most $q_j \eta$ over X_1^n, we can find α and γ such that

$$E_{\mathbf{X}} f_1(\mathcal{L}_j(\alpha, \mathbf{x})) > E_{X_1^n} f_2^{\gamma_i}(\mathcal{L}_j(\alpha, \mathbf{x})) + \sqrt{\frac{32}{n}\left(\ln 4\mathcal{N}_{\infty}(\mathcal{L}_j, \gamma_i, X_1^n) + \ln \frac{1}{p_i q_j \eta}\right)}.$$

Summing the probability over j, we obtain the corollary. ■

If close to perfect generalization can be achieved, i.e. $E_{X_1^n} f_2^{\gamma}(\mathcal{L}(\alpha, \mathbf{x})) \approx 0$, we can obtain better bounds by using a refined version of the Chernoff bound where the quantity $-2n\epsilon^2$ on the exponent can be replaced by $-n\epsilon^2/2(Ef + \epsilon)$ if the empirical mean is larger than the true mean; and by $-n\epsilon^2/Ef$ if the empirical mean is smaller than the true mean. In the extreme case that there is always a choice of α that achieves the perfect generalization: $E_{\mathbf{X}} f_2^{\gamma}(\mathcal{L}(\alpha, \mathbf{x})) = 0$, we can assume that our choice of $\alpha(X_1^n)$ satisfies $E_{X_1^n} f_2^{\gamma}(\mathcal{L}(\alpha, \mathbf{x})) = 0$. Under this assumption, bounds in this section can be improved substantially if we replace the standard Chernoff bound by the refined Chernoff bound. Specifically, a PAC bound in the order of $O\left(\frac{1}{n} \log \mathcal{N}_{\infty}\right)$ can be obtained, rather than a bound in the order of $O\left(\sqrt{\frac{1}{n} \log \mathcal{N}_{\infty}}\right)$ as in the standard case (3).

3. Covering Number Bounds for Linear Function Classes

Theorems in Section 2 indicate that covering numbers of a function class play crucial roles in the uniform convergence behavior of its members' empirical risks to their corresponding true risks. A bound on the rate of uniform convergence directly implies a bound on the generalization ability of an empirical risk minimization algorithm.

In this section, we derive some new covering number results for real valued linear function classes of the following form:

$$L(\mathbf{w}, \mathbf{x}) = \mathbf{w} \cdot \mathbf{x} = \sum_{j=1}^{d} \mathbf{x}^j \mathbf{w}^j. \tag{4}$$

We use $\mathbf{x}^j$ to denote the j-th component of the observation vector $\mathbf{x}$. We also use $\mathbf{w}$ to denote the linear weight, and $\mathbf{w}^j$ to denote its j-th component. d is the dimension of the system.

Results in this section complement related results in many previous studies such as those by Bartlett (1998), Guo et al. (1999), Mendelson (2001), Shawe-Taylor et al. (1998), Williamson et al. (1999, 2000). From theorems in Section 2, we see that covering number bounds for (4) are relevant to the learning behavior of (1). However, keep in mind that in this Section, notation L in (4) is used to denote a linear function class. This should not be confused with $\mathcal{L}$ in (1), which is used to denote a specific loss function in a specific learning formulation.

Covering number results of a linear function class such as those we derive in this section can also be used to derive covering numbers for certain nonlinear function classes. For example, if we can write a loss function as $\mathcal{L}(w, x, y) = \rho(f(w, x), y)$ where ρ is a Lipschitz function, then covering number bounds of $\mathcal{L}$ can be obtained using covering numbers of $f(\cdot)$ (see Lemma 17.6 of Anthony and Bartlett, 1999). Using this result, it is clear that bounds derived in this section can also be used to obtain covering numbers for more complicated functions such as neural networks and support vector machines.

We start with covering number results for Theorem 1. Because $\mathcal{N}_1 \leq \mathcal{N}_2$, therefore in order to apply Theorem 1, it is sufficient to estimate $\mathcal{N}_2(L, \epsilon, n)$ for $\epsilon > 0$. It is clear that $\mathcal{N}_2(L, \epsilon, n)$ is not finite if no restrictions on $\mathbf{x}$ or $\mathbf{w}$ are imposed. Therefore in the following, we assume the condition that $\|\mathbf{x}_i\|_p$ is bounded for observed data. We then focus on the form of regularization conditions on $\|\mathbf{w}\|_q$, so that $\log \mathcal{N}(f, \epsilon, n)$ is independent (or weakly dependent) of d.

We start our analysis with a lemma that is attributed to Maurey (also see Barron, 1993, Jones, 1992).

Lemma 1 (Maurey) *In a Hilbert space, let $f = \sum_{j=1}^{d} w_j \mathbf{g}^j$, where each $\|\mathbf{g}^j\| \leq b$, $w_j \geq 0$ and $\alpha = \sum_{j=1}^{d} w_j \leq 1$, then for every $n \geq 1$, there exist non-negative integers $k_1, \ldots, k_d \geq 0$, such that $\sum_{j=1}^{d} k_j \leq n$ and*

$$\left\| f - \frac{1}{n} \sum_{j=1}^{d} k_j \mathbf{g}^j \right\|^2 \leq \frac{\alpha b^2 - \|f\|^2}{n}.$$

Our first result generalizes a theorem of Bartlett (1998). The original result was with $p = \infty$ and $q = 1$; some related techniques have also been used by Lee et al. (1996) and Schapire et al. (1998). We would like to mention that it is possible to prove a better bound (when $p < \infty$) using machineries from the geometric theory of Banach spaces. However, the technique presented here is more elementary and self-contained. It is directly comparable to the idea of using online mistake bounds to obtain ∞-norm covering number bounds which we shall investigate later. This comparison provides useful insights.

Theorem 3 *If $\|\mathbf{x}\|_p \leq b$, and $\|\mathbf{w}\|_q \leq a$, where $1/p + 1/q = 1$ and $2 \leq p \leq \infty$, then*

$$\log_2 \mathcal{N}_2(L, \epsilon, n) \leq \left\lceil \frac{a^2 b^2}{\epsilon^2} \right\rceil \log_2(2d + 1).$$

Proof Consider matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, where each $\mathbf{x}_i$ is a column vector and T is the matrix transpose operator. Denote the columns of X by $\mathbf{y}^1, \ldots, \mathbf{y}^d$. Let

$$\mathbf{g}^j = \frac{n^{1/p}ab}{\|\mathbf{y}^j\|_p}\mathbf{y}^j, \qquad w'_j = \frac{\|\mathbf{y}^j\|_p}{n^{1/p}ab}\mathbf{w}^j,$$

where $\mathbf{w}^j$ is the j-th component of vector $\mathbf{w}$. By Hölder's inequality, it is easy to check that

$$\begin{aligned}\sum_{j=1}^d |w'_j| &= \sum_{j=1}^d \frac{\|\mathbf{y}^j\|_p}{n^{1/p}ab}|\mathbf{w}^j| \\ &\leq \frac{1}{n^{1/p}ab}\left(\sum_{j=1}^d \|\mathbf{y}^j\|_p^p\right)^{1/p}\left(\sum_{j=1}^d |\mathbf{w}^j|^q\right)^{1/q} \\ &\leq \frac{1}{n^{1/p}ab}(nb^p)^{1/p}a = 1.\end{aligned}$$

Since the function $x^{p/2}$ is convex, thus by Jensen's inequality, we obtain $n^{-1/2}\|\mathbf{y}^j\|_2 \leq n^{-1/p}\|\mathbf{y}^j\|_p$ for all j. This implies that $\|\mathbf{g}^j\|_2 \leq n^{1/2}ab$. Therefore by Lemma 1, if we let $k \geq (ab/\epsilon)^2$, then $\forall \mathbf{z} = \sum_{j=1}^d \mathbf{w}^j\mathbf{y}^j = \sum_{j=1}^d |w'_j|(\mathrm{sgn}(w'_j)\mathbf{g}^j)$, we can find integers $k_1, \ldots, k_d$ such that $\sum_{j=1}^d |k_j| \leq k$ and

$$\left\|\mathbf{z} - \frac{1}{k}\sum_{j=1}^d k_j\mathbf{g}^j\right\|_2^2 \leq \frac{na^2b^2}{k} \leq n\epsilon^2.$$

This means that the covering number $\mathcal{N}_2(L, \epsilon, n)$ is no larger than the number of integer solutions of $\sum_{j=1}^d |k_j| \leq k$, which is less than or equal to $(2d+1)^k$. ■

In the above proof, a more careful analysis of the number of possible solutions of $\sum_{j=1}^d |k_j| \leq k$ can lead to a bound of $(\frac{2e(d+k)}{k})^k$. Although when k is large this bound is tighter than the bound $(2d+1)^k$ used in Theorem 3, the difference is relatively minor since the dominant contribution is the exponent k. We have thus chosen the more compact expression $(2d+1)^k$.

The above bound on the (logarithmic) covering number depends logarithmically on d, which is already quite weak (compared to the linear d-dependency in the standard situation without regularization). However, it is also possible to remove the dimensional dependency. We demonstrate for the case of $p = 2$ in the following corollary.

Corollary 3 *If $\|\mathbf{x}\|_2 \leq b$ and $\|\mathbf{w}\|_2 \leq a$, then*

$$\log_2 \mathcal{N}_2(L, \epsilon, n) \leq \left\lceil \frac{a^2b^2}{\epsilon^2} \right\rceil \log_2(2n+1).$$

Proof For a sequence of data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, denote by S the subspace spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Denote by $P_S(\mathbf{w})$ the orthonormal projection operator that projects $\mathbf{w}$ onto the subspace S. Clearly, $P_S(\mathbf{w}) \cdot \mathbf{x}_i = \mathbf{w} \cdot \mathbf{x}_i$ for all $i = 1, \ldots, n$. This means that a data-dependent cover of $L(P_S(\mathbf{w}), \mathbf{x})$ also gives a data-dependent cover of $L(\mathbf{w}, \mathbf{x})$ in (4). To bound the 2-norm covering number of $L(P_S(\mathbf{w}), \mathbf{x})$, we simply observe that $\|P_S(\mathbf{w})\|_2 \leq \|\mathbf{w}\|_2 \leq a$ and that S is of dimension at most n. We thus obtain the corollary from Theorem 3. ■

Intuitively, the reason that we can remove the dimensional dependency of the 2-norm covering number in Corollary 3 is based on the observation that (in the case of $p = 2$) the effective dimension of $\mathbf{w}$, acted on n data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, is at most n. Using the same underlying idea, it is also possible to obtain dimension independent covering number results for the case $2 \leq p < \infty$. In general, for regularization condition $g(w) \leq a$, one needs to cover $\mathbf{w}$ with weights in the form of $\nabla g^{-1}(\sum_{i=1}^n \alpha_i \mathbf{x}_i)$, where ∇g^{-1} is the inverse function of the gradient ∇g. This changes the dimension d of $\mathbf{w}$ to the dimension n of α.

Naturally the above discussion leads to the idea of using the mistake bound framework in online learning to obtain covering number bounds. In online learning, one represents the weight vector as $\nabla g^{-1}(\sum_{i=1}^n \alpha_i \mathbf{x}_i)$, where the function ∇g^{-1} is called a transfer function (Grove et al., 2001, Gentile and Warmuth, 1998, Kivinen and Warmuth, 1997). An online mistake bound can be regarded as an approximation bound for this representation. At each step, we look at a data point $\mathbf{x}_i$, and check the prediction (such as classification) associated with the current weight representation. We add $\mathbf{x}_i$ into the representation only when a mistake is made. The mistake bound analysis by Grove et al. (2001) shows that for certain classification problems, there exists a quantity M so that after M components are added in the representation, no more mistakes will be made (thus no more components will be added). In this regard, the sparse representation in Maurey's lemma (Lemma 1) corresponds to the sparse representation in the mistake bound framework (Grove et al., 2001). We can thus use the latter to upper bound the sparsity k in the representation $\nabla g^{-1}(\sum_{j=1}^k \eta \mathbf{x}_{i_j})$ such that $|\nabla g^{-1}(\sum_{j=1}^k \eta \mathbf{x}_{i_j}) - \mathbf{w} \cdot \mathbf{x}_i| \leq \epsilon$ for all i with any specified $\epsilon > 0$. This idea is rigorously carried out in the proof of Theorem 4 below.

Also note that using online learning, we are able to directly obtain bounds for ∞-norm covering numbers, which are not only useful for Theorem 1 (since $\mathcal{N}_1 \leq \mathcal{N}_\infty$), but also useful for Theorem 2. Therefore we shall not further consider the idea of using Maurey's lemma and restrict the effective dimension to remove the d-dependency as in Corollary 3. We will only focus on using online learning techniques to directly obtain ∞-norm covering numbers.

We shall mention that traditionally, ∞-norm covering numbers are obtained through the so-called "fat-shattering" dimension (Alon et al., 1997). The latter can be bounded using various methods (Bartlett, 1998, Gurvits, 1997, Shawe-Taylor et al., 1998). However, due to the extra stage of estimating the "fat-shattering" dimension, this approach leads to bounds that are worse than our bounds that are directly obtained. For example, the "fat-shattering" approach would have led to a bound of the order $\log_2 \mathcal{N}_\infty(L, \epsilon, n) = O(\frac{a^2 b^2}{\epsilon^2} \log_2(n/\epsilon + 1)^2)$ for $p = 2$ in Theorem 4. Since covering numbers (rather than fat-shattering dimensions) have more direct learning consequences, our results can lead to better learning bounds than what can be obtained from these earlier studies (see Section 4).

In the proof of Theorem 4, we need the following mistake bound result that was proved by Grove et al. (2001). The bound indicates that we can approximate a target vector $\mathbf{w}$ (that satisfies a margin property) using no more than M data points, so that the inner product of the approximation vector and any data point x_i has the same sign as that of $\mathbf{w} \cdot \mathbf{x}_i$.

Proposition 1 *Consider $2 \leq p < \infty$. Let $\mathbf{w}$ be a target vector, and $\{\mathbf{x}_i\}$ be a countable set of example vectors. Assume that $\delta = \inf_i \mathbf{w} \cdot \mathbf{x}_i > 0$, and let*

$$M = \frac{(p-1)\|\mathbf{w}\|_q^2 \sup_i \|\mathbf{x}_i\|_p^2}{\delta^2}.$$

Then there exists an integer sequence $i_1, \ldots, i_k$ where $k \leq M$, and a vector $\hat{\mathbf{w}}$ defined as $\hat{\mathbf{w}} = f_p(\sum_{\ell=1}^{k} \mathbf{x}_{i_\ell})$ so that $\hat{\mathbf{w}} \cdot \mathbf{x}_i > 0$ for all i. $f_p(\mathbf{x})$ is a component-wise function that maps each component $\mathbf{x}^j$ of $\mathbf{x}$ to $p \cdot \mathrm{sign}(\mathbf{x}^j)|\mathbf{x}^j|^{p-1}$.

Theorem 4 *If $\|\mathbf{x}\|_p \leq b$ and $\|\mathbf{w}\|_q \leq a$, where $2 \leq p < \infty$ and $1/p + 1/q = 1$, then $\forall \epsilon > 0$,*

$$\log_2 \mathcal{N}_\infty(L, \epsilon, n) \leq 36(p-1)\frac{a^2 b^2}{\epsilon^2} \log_2[2\lceil 4ab/\epsilon + 2\rceil n + 1].$$

Proof If $\epsilon > ab$, then since $|\mathbf{w} \cdot \mathbf{x}_i| \leq ab$ for all i, we can choose 0 as a cover and the theorem follows trivially. In the following we assume that $\epsilon \leq ab$.

We divide the interval $[-ab - \epsilon/2, ab + \epsilon/2]$ into $m = \lceil 4ab/\epsilon + 2\rceil$ sub-intervals, each of size no larger than $\epsilon/2$. Let $-ab - \epsilon/2 = \theta_0 < \theta_1 < \cdots < \theta_m = ab + \epsilon/2$ be the boundaries of the intervals so that $\theta_j - \theta_{j-1} \leq \epsilon/2$ for all j. For a sample $X_1^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, consider the sets $S_1 = \{(\mathbf{x}_i, -\theta_j/a) : i = 1, \ldots, n; j = 0, \ldots, m-1\}$ and $S_2 = \{(-\mathbf{x}_i, \theta_j/a) : i = 1, \ldots, n; j = 1, \ldots, m\}$.

For all $\mathbf{w}$ such that $\|\mathbf{w}\|_q \leq a$, consider the set of values $\mathbf{w} \cdot \mathbf{x}_i - \theta_{j_1(i,\mathbf{w})}$ and $-\mathbf{w} \cdot \mathbf{x}_i + \theta_{j_2(i,\mathbf{w})}$ for all i. We use $j_1(i, \mathbf{w})$ to denote the maximum index of θ_j such that $\mathbf{w} \cdot \mathbf{x}_i - \theta_{j_1(i,\mathbf{w})} \geq \epsilon/2$; and use $j_2(i, \mathbf{w})$ to denote the minimum index of θ_j such that $\mathbf{w} \cdot \mathbf{x}_i - \theta_{j_2(i,\mathbf{w})} \leq -\epsilon/2$.

Now, we consider $(\mathbf{y}, z)$ such that $\forall i$: $\mathbf{y} \cdot \mathbf{x}_i - z\theta_{j_1(i,\mathbf{w})} > 0$ and $-\mathbf{y} \cdot \mathbf{x}_i + z\theta_{j_2(i,\mathbf{w})} > 0$. Since $\theta_{j_1(i,\mathbf{w})} < \theta_{j_2(i,\mathbf{w})}$, it follows that $z > 0$ and $\forall i : \mathbf{y} \cdot \mathbf{x}_i / z \in (\theta_{j_1(i,\mathbf{w})}, \theta_{j_2(i,\mathbf{w})})$. This implies that $|\mathbf{y} \cdot \mathbf{x}_i / z - \mathbf{w} \cdot \mathbf{x}_i| < \epsilon$ for all i.

Next we show how to construct such a pair of $(\mathbf{y}, z)$. Let $f_p(z) = p \cdot \mathrm{sign}(z)|z|^{p-1}$, and

$$M = 36(p-1)\frac{a^2 b^2}{\epsilon^2} \geq \frac{(p-1)}{(\epsilon/2)^2}(\|\mathbf{w}\|_q^q + a^q)^{2/q} \sup_i (\|\mathbf{x}_i\|_p^p + (b + \epsilon/2a)^p)^{2/p}.$$

Using Proposition 1, we know that $\forall \|\mathbf{w}\|_q \leq a$, there exist non-negative integer sequences α_i and β_i where $\sum_{i=1}^{n}(\alpha_i + \beta_i) \leq M$, with the following property: if we let

$$(\mathbf{y}, az) = f_p\left(\sum_i \alpha_i(\mathbf{x}_i, -\theta_{j_1(i,\mathbf{w})}/a) + \sum_i \beta_i(-\mathbf{x}_i, \theta_{j_2(i,\mathbf{w})}/a)\right),$$

then $\mathbf{y} \cdot \mathbf{x}_i - z\theta_{j_1(i,\mathbf{w})} > 0$ and $-\mathbf{y} \cdot \mathbf{x}_i + z\theta_{j_2(i,\mathbf{w})} > 0$ for all i.

It follows from the above discussion that $|\mathbf{y} \cdot \mathbf{x}_i/z - \mathbf{w} \cdot \mathbf{x}_i| < \epsilon$ for all i. This implies that the infinity-norm covering number $\mathcal{N}_\infty(L, \epsilon, n)$ is no more than the number of possible $(\mathbf{y}, z)$ constructed above. It is clear that this number is no more than the number of non-negative integer solutions of

$$\sum_{i,j} n_{i,j} + \sum_{i,j} m_{i,j} \leq M,$$

where (i, j) goes through the index of S_1 for $n_{i,j}$ and the index of S_2 for $m_{i,j}$. Since the number of solutions is no more than $(|S_1| + |S_2| + 1)^M$, we obtain

$$\log_2 \mathcal{N}_\infty(L, \epsilon, n) \leq 36(p-1)\frac{a^2 b^2}{\epsilon^2} \log_2[2\lceil 4ab/\epsilon + 2\rceil n + 1].$$

■

The bound given in Theorem 1 is comparable to related results by Williamson et al. (2000), which were obtained by using a different technique relying on the operator theory of Banach spaces. However, in certain cases, our method may yield results that are difficult to obtain from the Banach space approach considered by Williamson et al. (2000). For example, it is difficult to obtain the entropy regularization result in Theorem 5 using their method. One reason is that a topological structure is necessary for the argument of Williamson et al. (2000) to go through. As we shall see later, our analysis only involves some analytical structures of an appropriately defined pair of dual convex functions on the linear weight space and the sample space. On one hand, the norm of a Banach space naturally leads to a convex function on the underlying space; on the other hand, the concept of convex function can be defined on any linear vector space that does not necessarily have a norm or topological structure.

Note that we have made no attempt to optimize the constants in the proof of Theorem 4. For example, since $\theta_0 = -ab - \epsilon/2$ and $\theta_m = ab + \epsilon/2$ are quite artificially introduced for the mere purpose of consistent indexing, we can easily obtain an improved version of Theorem 4 by simply ignoring them. Also note that $\mathcal{N}_2 \leq \mathcal{N}_\infty$, therefore Theorem 4 implies dimension independent 2-norm covering number bounds for $2 \leq p < \infty$, which gives better results than Theorem 3 in the sense of dimensional dependency. The bound in Theorem 4 diverges as $p \to \infty$. In the case of $p = \infty$, we show that an entropy condition can be used to obtain dimension independent covering number bounds. This entropy condition is related to multiplicative update methods widely studied in online learning. To our knowledge, there are no previous covering number results for entropy regularization. We first introduce the following definition:

Definition 2 *Let $\mu = [\mu_j]$ be a vector with positive entries such that $\|\mu\|_1 = 1$ (in this case, we call μ a distribution vector). Let $\mathbf{x} = [\mathbf{x}^j] \neq 0$ be a vector of the same dimension, then we define the weighted relative entropy of $\mathbf{x}$ with respect to μ as:*

$$\mathrm{entro}_\mu(\mathbf{x}) = \sum_j |\mathbf{x}^j| \ln \frac{|\mathbf{x}^j|}{\mu_j \|\mathbf{x}\|_1}.$$

It is a well-known fact that the relative entropy defined above is always non-negative, and $\mathrm{entro}_\mu(\mathbf{x}) = 0$ only when $|\mathbf{x}| = \|\mathbf{x}\|_1 \cdot \mu$. Before the main theorem, we need a lemma that refines and generalizes the analysis of Grove et al. (2001, Section 6). Note that for our purpose, their result is not directly applicable. Related techniques have also been used by Gentile and Warmuth (1998), Kivinen and Warmuth (1997). In the following lemma, $\mathbf{x}_i^j$ indicates the j-th component of vector $\mathbf{x}_i$.

Lemma 2 *Let μ be a distribution vector and $\mathbf{w}$ be a vector with non-negative entries such that $\|\mathbf{w}\|_1 \leq W$. Assume that $\{\mathbf{x}_i\}$ is a countable set of example vectors such that $\inf_i \mathbf{w}\cdot\mathbf{x}_i > 0$. $\forall \delta \in (0, \min_i \mathbf{w}\cdot\mathbf{x}_i]$, let*

$$m(\delta) = \frac{2\sup_i \|\mathbf{x}_i\|_\infty^2 W \cdot \mathrm{entro}_\mu(\mathbf{w})}{\delta^2}.$$

Then there exists an integer sequence $i_1, \ldots, i_k$ where $k \leq m(\delta)$, and a vector $\hat{\mathbf{w}}$ with its j-th component defined as $\hat{\mathbf{w}}^j = \mu_j \exp(\eta \sum_{\ell=1}^k \mathbf{x}_{i_\ell}^j)$, where $\eta = \delta/(W \sup_i \|\mathbf{x}_i\|_\infty^2)$, so that $\hat{\mathbf{w}}\cdot\mathbf{x}_i > 0$ for all i.

Proof Without loss of generality, we assume that $\|\mathbf{w}\|_1 = 1$. Let $\mathbf{z}$ be a vector, and consider the convex dual of $\sum_{j=1}^d \mathbf{w}^j \ln \frac{\mathbf{w}^j}{\mu_j}$:

$$\ln \sum_{j=1}^d \mu_j e^{\mathbf{z}^j} = \sup_{\|\mathbf{w}\|_1=1} \left[\mathbf{w}\cdot\mathbf{z} - \sum_{j=1}^d \mathbf{w}^j \ln \frac{\mathbf{w}^j}{\mu_j} \right].$$

The definition of this duality implies that the quantity

$$M(\mathbf{z}) = \ln \sum_{j=1}^d \mu_j e^{\mathbf{z}^j} - \mathbf{w}\cdot\mathbf{z} + \sum_{j=1}^d \mathbf{w}^j \ln \frac{\mathbf{w}^j}{\mu_j}$$

is always non-negative.

Assume now that the theorem is not true, then there exists a sequence of integers $i_1, \ldots, i_k$ where $k > m(\delta)$ such that if we define a sequence of vectors $\mathbf{z}_\ell$ as $\mathbf{z}_\ell = \mathbf{z}_{\ell-1} + \eta \mathbf{x}_{i_\ell}$ with $\mathbf{z}_0 = 0$, then $\sum_{j=1}^d \mu_j \exp(\mathbf{z}_{\ell-1}^j)\mathbf{x}_{i_\ell}^j \leq 0$ for $\ell = 1, 2, \ldots, k$.

Note that for all pairs of vectors $(\mathbf{v}, \Delta\mathbf{v})$:

$$\frac{d}{dt} \ln \sum_{j=1}^d \mu_j e^{\mathbf{v}^j + \Delta\mathbf{v}^j t} = \frac{\sum_{j=1}^d \mu_j e^{\mathbf{v}^j + \Delta\mathbf{v}^j t} \Delta\mathbf{v}^j}{\sum_{j=1}^d \mu_j e^{\mathbf{v}^j + \Delta\mathbf{v}^j t}}$$

and

$$\frac{d^2}{dt^2} \ln \sum_{j=1}^d \mu_j e^{\mathbf{v}^j + \Delta\mathbf{v}^j t} \leq \frac{\sum_{j=1}^d \mu_j e^{\mathbf{v}^j + \Delta\mathbf{v}^j t} \Delta\mathbf{v}^{j\,2}}{\sum_j \mu_j e^{\mathbf{v}^j + \Delta\mathbf{v}^j t}}.$$

Therefore by Taylor expansion, we know that there exists $t \in [0,1]$ such that

$$\begin{aligned}
\ln \sum_{j=1}^{d} \mu_j e^{\mathbf{z}_\ell^j} &\leq \ln \sum_{j=1}^{d} \mu_j e^{\mathbf{z}_{\ell-1}^j} + \frac{\sum_{j=1}^d \mu_j e^{\mathbf{z}_{\ell-1}^j} \eta \mathbf{x}_{i_\ell}^j}{\sum_{j=1}^d \mu_j e^{\mathbf{z}_{\ell-1}^j}} + \frac{\eta^2}{2} \frac{\sum_{j=1}^d \mu_j e^{\mathbf{z}_{\ell-1}^j + \eta \mathbf{x}_{i_\ell}^j t} \mathbf{x}_{i_\ell}^{j\,2}}{\sum_{j=1}^d \mu_j e^{\mathbf{z}_{\ell-1}^j + \eta \mathbf{x}_{i_\ell}^j t}} \\
&\leq \ln \sum_{j=1}^{d} \mu_j e^{\mathbf{z}_{\ell-1}^j} + \frac{\eta^2}{2} \frac{\sum_{j=1}^d \mu_j e^{\mathbf{z}_{\ell-1}^j + \eta \mathbf{x}_{i_\ell}^j t} \mathbf{x}_{i_\ell}^{j\,2}}{\sum_{j=1}^d \mu_j e^{\mathbf{z}_{\ell-1}^j + \eta \mathbf{x}_{i_\ell}^j t}} \\
&\leq \ln \sum_{j=1}^{d} \mu_j e^{\mathbf{z}_{\ell-1}^j} + \frac{\eta^2}{2} \|\mathbf{x}_{i_\ell}\|_\infty^2.
\end{aligned}$$

Note that we have used the assumption that $\sum_{j=1}^d \mu_j \exp(\mathbf{z}_{\ell-1}^j) \mathbf{x}_{i_\ell}^j \leq 0$ in the above derivation. We now obtain

$$\begin{aligned}
M(\mathbf{z}_\ell) - M(\mathbf{z}_{\ell-1}) &= \ln \frac{\sum_{j=1}^d \mu_j e^{\mathbf{z}_\ell^j}}{\sum_{j=1}^d \mu_j e^{\mathbf{z}_{\ell-1}^j}} - \mathbf{w} \cdot \eta \mathbf{x}_{i_\ell} \\
&\leq \frac{\eta^2}{2} \|\mathbf{x}_{i_\ell}\|_\infty^2 - \eta\delta.
\end{aligned}$$

Summing the above inequality over ℓ, and note that $W \geq \|\mathbf{w}\|_1 = 1$:

$$\begin{aligned}
M(\mathbf{z}_k) <& M(\mathbf{z}_0) + m(\delta)(\frac{\eta^2}{2} \sup_i \|\mathbf{x}_i\|_\infty^2 - \eta\delta) \\
=& \text{entro}_\mu(\mathbf{w}) + m(\delta)(\frac{\eta^2}{2} \sup_i \|\mathbf{x}_i\|_\infty^2 - \eta\delta) \leq 0,
\end{aligned}$$

which is a contradiction since $M(\mathbf{z}_k)$ is always non-negative. ■

Theorem 5 *Given a distribution vector μ, if $\|\mathbf{x}\|_\infty \leq b$ and $\|\mathbf{w}\|_1 \leq a$ and $\text{entro}_\mu(\mathbf{w}) \leq c$, where we assume that $\mathbf{w}$ has non-negative entries, then $\forall \epsilon > 0$,*

$$\log_2 \mathcal{N}_\infty(L, \epsilon, n) \leq \frac{36b^2(a^2 + ac)}{\epsilon^2} \log_2[2\lceil 4ab/\epsilon + 2 \rceil n + 1].$$

Proof The proof follows the same steps of Theorem 4. We let $\mu' = [\mu, 1]/2$ and $\mathbf{w}' = [\mathbf{w}, a]$. We have $\|\mathbf{w}'\|_1 \leq 2a$, and $\text{entro}_{\mu'}(\mathbf{w}') \leq \text{entro}_\mu(\mathbf{w}) + a \ln 2 < a + c$. Similarly, the expansion $\mathbf{x}_i'$ of $\mathbf{x}_i$ (by appending a component θ/a) satisfies $\|\mathbf{x}_i'\|_\infty \leq 1.5b$ (again, we assume that $\epsilon/a \leq b$).

We now apply the mistake bound in Lemma 2, where we set $\delta = \epsilon/2$ and $W = 2a$. We can define M as

$$M = \frac{36(a+c)ab^2}{\epsilon^2} \geq \frac{2}{\delta^2} \sup_i \|\mathbf{x}_i'\|_\infty^2 W \cdot \text{entro}_{\mu'}(\mathbf{w}').$$

The remaining part of the proof is the same as that of Theorem 4. ■

Corollary 4 *Given a distribution vector μ, if $\|\mathbf{x}\|_\infty \le b$ and $\|\mathbf{w}\|_1 \le a$ and $\mathrm{entro}_\mu(\mathbf{w}) \le c$, then $\forall \epsilon > 0$,*

$$\log_2 \mathcal{N}_\infty(L, \epsilon, n) \le \frac{288b^2(2a^2 + ac)}{\epsilon^2} \log_2[2\lceil 8ab/\epsilon + 2\rceil n + 1].$$

Proof Define vector $\mathbf{u}$ component-wise as $\max(\mathbf{w}, 0)$, and similarly define $\mathbf{v} = \max(-\mathbf{w}, 0)$. By definition, $\mathbf{w} = \mathbf{u} - \mathbf{v}$ and $\|\mathbf{u}\|_1, \|\mathbf{v}\|_1 \le \|\mathbf{w}\|_1$. For all $L = L_1 - L_2$, we have $\mathcal{N}_\infty(L, \epsilon, n) \le \mathcal{N}_\infty(L_1, \epsilon/2, n) \cdot \mathcal{N}_\infty(L_2, \epsilon/2, n)$. Therefore we only need to show that $\mathrm{entro}_\mu(\mathbf{u}) \le \mathrm{entro}_\mu(\mathbf{w}) + \|\mathbf{w}\|_1$. To prove this, we shall assume that $\|\mathbf{w}\|_1 = 1$ without loss of generality, and $\mathbf{u}, \mathbf{v} \ne 0$. Since $\|\mathbf{u}\|_1 + \|\mathbf{v}\|_1 = 1$,

$$\|\mathbf{u}\|_1 \ln \frac{1}{\|\mathbf{u}\|_1} + \|\mathbf{v}\|_1 \ln \frac{1}{\|\mathbf{v}\|_1} \le \ln 2 \le \ln 2 + \sum_{j=1}^{d} \mathbf{v}^j \ln \frac{\mathbf{v}^j}{\|\mathbf{v}\|_1 \mu_j}.$$

The above inequality can be rewritten as

$$\sum_{j=1}^{d} \mathbf{u}^j \ln \frac{\mathbf{u}^j}{\mu_j \|\mathbf{u}\|_1} \le \ln 2 + \sum_{j=1}^{d} \mathbf{u}^j \ln \frac{\mathbf{u}^j}{\mu_j} + \sum_{j=1}^{d} \mathbf{v}^j \ln \frac{\mathbf{v}^j}{\mu_j}.$$

That is, $\mathrm{entro}_\mu(\mathbf{u}) \le \mathrm{entro}_\mu(\mathbf{w}) + \ln 2$. ■

Note that we don't require the dimension to be finite. However, assume that the dimension d is finite, and we let $\mu_j = 1/d$. Then it is easy to check that $\forall \mathbf{w}, \mathrm{entro}_\mu(\mathbf{w}) \le \|\mathbf{w}\|_1 \ln d$. Therefore by Corollary 4, we obtain the following result which gives a better bound than a similar result of Bartlett (1998) by a logarithmic factor of n.

Corollary 5 *If $\|\mathbf{x}\|_\infty \le b$ and $\|\mathbf{w}\|_1 \le a$, then $\forall \epsilon > 0$,*

$$\log_2 \mathcal{N}_\infty(L, \epsilon, n) \le \frac{288a^2b^2(2 + \ln d)}{\epsilon^2} \log_2[2\lceil 8ab/\epsilon + 2\rceil n + 1].$$

We now discuss the relationship among different covering number bounds obtained in this section. Theorem 3 uses a reduction technique to generalize a result of Bartlett (1998). The derivation employs Maurey's Lemma. By observing that the effective dimension is no larger than n, it is possible to remove the inherent logarithmic dependency on dimension d for certain regularization conditions, as demonstrated in Corollary 3.

Using online learning, this idea can be more systematically developed. For example, Theorem 4 (note that $\mathcal{N}_2 \le \mathcal{N}_\infty$) employs the online mistake bound framework, which leads to a bound with the $\log d$ dependency replaced by a $\log n$ dependency. This trade-off of $\log d$ and $\log n$ is very natural from the computational point of view since Maurey's Lemma achieves an approximation by selecting columns (relevant features) of the data while an online algorithm achieves an approximation by selecting rows (related to support vectors) of the data.

It follows that if $d \ll n$, then Theorem 4 gives a better bound; and if $n \ll d$, then Theorem 3 gives a better bound. However, if we use these covering number results in a PAC style generalization analysis, then a $\log n$ dependency on the sample size usually does not cause any substantial problem.

In the proof of Corollary 3 the effective dimension of the problem is reduced by a compactification of part of the dimensions. To a certain extent, all dimension independent covering number results obtained in this section implicitly rely on the (weak) compactness of the effective parameter family in one way or another. Theorem 5 achieves this through the entropy regularization condition. If we regard μ_j as a prior measure and $\mathbf{w}$ as a posterior measure, then the entropy condition in Theorem 5 corresponds to the maximum entropy principle in density estimation, which can be regarded as entropy regularization. Therefore the dimension independent covering number result justifies the maximum entropy method from the PAC learning point of view.

If we let $p \to \infty$, then the covering number bound given in Theorem 4 diverges. It has been pointed out by Grove et al. (2001) that this divergence is a consequence of regularizing the weight parameter $\mathbf{w}$ around the origin. As a comparison, Theorem 5 gives a finite covering number for $p = \infty$ with entropy regularization. It is also possible to construct a regularization condition around a non-zero vector so that when $p \to \infty$, the bound in Theorem 4 approaches the limiting case of Theorem 5. Because of Theorem 5 and its relation to the well-established maximum entropy principle, it is reasonable to use entropy (instead of 1-norm) as the regularization condition for infinity-norm bounded data. For example, such a condition has recently been employed by Jaakkola et al. (2000). Entropy regularization has also been implicitly employed in the Winnow family of multiplicative update algorithms (Littlestone, 1988), and its continuous version of EG (and EGU) algorithms (Kivinen and Warmuth, 1997). In addition, Zhang (2002) used explicit entropy regularization conditions to convert EG online algorithms into batch learning algorithms.

In addition to the Maurey's lemma approach used in this paper, 2-norm covering number bounds can also be obtained by using an inequality from the theory of Gaussian processes, often referred to as Sudakov's minoration (see Ledoux and Talagrand, 1991, chapter 12). This inequality bounds the 2-norm covering number of a function class by the expectation of a Gaussian process indexed by the function class. The latter can be estimated, which some time leads to quite tight bounds. However we shall not include results obtained by this approach in this paper. There are also other methods to obtain p-norm covering number bounds, for example, by using the fat-shattering dimension of a function class (Mendelson, 2001).

We have shown in this section that infinity-norm covering number bounds can be derived from online mistake bounds. From the construction of $M(\mathbf{z})$ in the proof of Lemma 2, we see that weight $\mathbf{w}$ and data $\mathbf{z}$ are Legendre dual variables with respect to the regularization condition g (as well as its convex dual function). The representation of $\mathbf{z}$ as a linear combination of data $\mathbf{x}_i$ leads to a dual representation of $\mathbf{w}$. This Legendre duality transforms the learning problem from the original d-dimensional space (where $\mathbf{w}$ is represented by its components) into the n-dimensional dual space (where $\mathbf{z}$ is represented by a linear combination of the data). This is why the logarithmic dimension factor $\log d$ in Maurey's Lemma (in the original space) can be replaced by $\log n$ in the dual (online learning) approach. The basic idea of reducing the effective dimension of $\mathbf{w}$ by using a linear combination in the

sample space has also appeared in the derivation of "fat-shattering" dimensions by Gurvits (1997), although an entirely different approach was employed there. However, the online learning approach used here that utilizes convex duality is more general. For example, we are able to derive covering number bounds for entropy regularization, which cannot be handled by previous techniques. Furthermore, this convex duality can be easily generalized so that given any convex potential on $\mathbf{x}$, we can obtain covering number bounds with the corresponding dual regularization condition on $\mathbf{w}$. We would also like to mention that if formulation (1) is convex, then we can study the dual representation of this formulation directly (Zhang, 2002). Such a dual representation is closely related to the online learning duality employed in this paper.

Finally, it is interesting to observe that the technique used in the proof of Lemma 2 is closely related to the potential-reduction method for linear programming (Todd, 1997, and references therein), where a variant of $M(\mathbf{z})$ with a flipped sign for the second term can be used to show the polynomial convergence of certain interior point algorithms. Similar to the proof of Lemma 2, the technique of bounding the number of steps is also based on a constant reduction of the potential function at each step, which is achieved by choosing an appropriate η from the first and the second order terms in a Taylor expansion. However, since Newton steps are often taken, the techniques required for bounding such terms are more complicated.

4. Some Consequences of Covering Number Bounds

In this section, we illustrate some simple consequences of our covering number bounds on some specific learning formulations. There are a number of books that describe how to use covering numbers to analyze learning algorithms (Anthony and Bartlett, 1999, Cristianini and Shawe-Taylor, 2000, Vapnik, 1998). Together with our covering number results for regularized linear function classes, machineries developed there can be used to study the generalization behavior of linear learning formulations such as (1).

For simplicity, we only include some direct consequences in this section. As an example, we can easily obtain the following bound from Theorem 4, which improves a related result of Bartlett and Shawe-Taylor (1999), and Cristianini and Shawe-Taylor (2000).

Theorem 6 *If the data is 2-norm bounded as $\|\mathbf{x}\|_2 \leq b$, then consider the family Γ of hyperplanes $\mathbf{w}$ such that $\|\mathbf{w}\|_2 \leq a$. Denote by $err(\mathbf{w})$ the misclassification error of $\mathbf{w}$ with the true distribution. Then there is a constant C such that with probability $1-\eta$ over $n > 1$ random samples, for all $\gamma > 0$ and $w \in \Gamma$ we have*

$$err(\mathbf{w}) \leq \frac{k_\gamma}{n} + \sqrt{\frac{C}{n}\left(\frac{a^2b^2}{\gamma^2}\ln(n) + \ln\frac{1}{\eta}\right)},$$

where $k_\gamma = |\{i : w^T x^i y^i < \gamma\}|$ is the number of samples with margin less than γ.

Proof Using the covering number result in Theorem 4. Let $\gamma_i = ab/2^i$ and $p_i = 1/2^i = \gamma_i/(ab)$ in Corollary 1. We have the bound:

$$err(\mathbf{w}) \leq \frac{k_\gamma}{n} + O\left(\sqrt{\frac{1}{n}\left(\frac{a^2b^2}{\gamma^2}\ln((\frac{ab}{\gamma}+1)n) + \ln\frac{ab}{\gamma} + \ln\frac{1}{\eta}\right)}\right),$$

Now using the fact that with an appropriate constant in the O notation, the above bound is trivial when $\gamma > ab$ or $\gamma n < ab$, it is easy to see that the above bound is equivalent to the claim of the theorem. ■

Note that the corresponding bound given by Cristianini and Shawe-Taylor (2000) was their Theorem 4.19. A similar theorem was also stated by Bartlett and Shawe-Taylor (1999). However in both cases, the $\ln n$ factor in our bound were replaced by $\ln^2 n$. The underlying technique leading to such a result came originally from Bartlett (1998), and fat-shattering dimension results by Anthony and Bartlett (1999), Shawe-Taylor et al. (1998). The reason why we can obtain a better bound in this paper is that the L_∞ covering number bound in Theorem 4 improves the corresponding bounds used previously, which were obtained through "fat-shattering" dimension estimates.

As another example, we consider the following bound:

Theorem 7 *If the data is infinity-norm bounded as $\|\mathbf{x}\|_\infty \leq b$, then consider the family Γ of hyperplanes $\mathbf{w}$ such that $\|\mathbf{w}\|_1 \leq a$. Let μ be a fixed non-negative prior vector. Denote by $err(\mathbf{w})$ the misclassification error of $\mathbf{w}$ with the true distribution. Then there is a constant C such that with probability $1-\eta$ over $n > 1$ random samples, for all γ and $\mathbf{w} \in \Gamma$, we have*

$$err(\mathbf{w}) \leq \frac{k_\gamma}{n} + \sqrt{\frac{C}{n}\left(\frac{b^2(a^2 + a\mathrm{entro}_\mu(\mathbf{w}))}{\gamma^2}\ln(n) + \ln\frac{1}{\eta}\right)},$$

where $k_\gamma = |\{i : w^T x^i y^i < \gamma\}|$ is the number of samples with margin less than γ.

Proof Consider the restriction of parameter $\mathbf{w}$ in $\Gamma_i \subseteq \Gamma$, where $\Gamma_1 = \{\mathbf{w} \in \Gamma : \mathrm{entro}_\mu(\mathbf{w}) \leq a\}$ and $\Gamma_j = \{\mathbf{w} \in \Gamma : \mathrm{entro}_\mu(\mathbf{w}) \in (2^{j-1}a, 2^j a]\}$ for $j > 1$. For each j, using Corollary 4 and essentially the same proof as that of Theorem 6, we obtain the following bound for the parameter family $\{\mathbf{w} \in \Gamma_j\}$:

$$err(\mathbf{w}) \leq \frac{k_\gamma}{n} + O\left(\sqrt{\frac{1}{n}\left(\frac{b^2(a^2 + a\mathrm{entro}_\mu(\mathbf{w}))}{\gamma^2}\ln(n) + \ln\frac{1}{\eta}\right)}\right),$$

Now let $q_j = 1/2^j$, which implies that $q_j = O(a/(a + \mathrm{entro}_\mu(\mathbf{w})))$. Using Corollary 2, we obtain the following bound for the parameter family $\{\mathbf{w} \in \Gamma\}$:

$$err(\mathbf{w}) \leq \frac{k_\gamma}{n} + O\left(\sqrt{\frac{1}{n}\left(\frac{a^2b^2c}{\gamma^2}\ln(n) + \ln c + \ln\frac{1}{\eta}\right)}\right),$$

where $c = (a + \mathrm{entro}_\mu(\mathbf{w}))/a$. It is now easy to see that the above bound leads to the theorem. ■

The above theorem is similar to a recent result of Langford and Seeger (2001) which was obtained by specialized techniques for PAC-Bayes analysis originally developed by McAllester (1999). If we let μ be the uniform prior, then the above bound easily leads to the following result of Schapire et al. (1998), which was used to explain the effectiveness of boosting.

Theorem 8 (Schapire et al. 1998) *If the data has dimension $d > 1$ and is infinity-norm bounded as $\|\mathbf{x}\|_\infty \leq 1$, then consider the family Γ of hyperplanes $\mathbf{w}$ with non-negative weights such that $\|\mathbf{w}\|_1 \leq 1$. Denote by $err(\mathbf{w})$ the misclassification error of $\mathbf{w}$ with the true distribution. Then there is a constant C with probability $1-\eta$ over $n > 1$ random samples, for all γ and $\mathbf{w} \in \Gamma$, we have*

$$err(\mathbf{w}) \leq \frac{k_\gamma}{n} + \sqrt{\frac{C}{n}\left(\frac{\ln(d)\ln(n)}{\gamma^2} + \ln\frac{1}{\eta}\right)},$$

where $k_\gamma = |\{i : w^T x^i y^i < \gamma\}|$ is the number of samples with margin less than γ.

It can be seen that our bound in Theorem 7 can be significantly better if one can guess a good prior μ so that $\mathrm{entro}_\mu(\hat{\mathbf{w}})$ is small for a parameter $\hat{\mathbf{w}}$ which has a good classification performance. If $\mathrm{entro}_\mu(\hat{\mathbf{w}})$ is small, then unlike Theorem 8, our bound can be dimension independent. This implies that it may be possible to vote an infinite number of classifiers so that the generalization performance is still good.

5. Discussion

In this paper, we have studied some theoretical aspects of using the regularization in linear learning formulations such as (1). We show that with appropriate regularization conditions, we can achieve the same dimension independent generalization performance enjoyed by support vector machines.

The separation concept introduced in Theorem 2 implies that the "margin" idea developed for linear classification can be naturally extended to general learning problems. Compared with Theorem 1, Theorem 2 is more suitable for problems with non-smooth loss functions since it does not directly employ the covering number of the overall loss function itself. Note that in general, the covering number of a function class depends on certain smoothness conditions of the family. Such smoothness requirement can lead to difficulties when we try to directly apply Theorem 1 to problems with non-smooth loss functions.

In Section 3, we have obtained some new covering number bounds for linear function classes under certain regularization conditions. These bounds have complemented and improved various previous results. We compared two different approaches for deriving covering number bounds. The first approach employs Maurey's lemma for sparsification. This approach has also been used in many previous studies of covering numbers (Anthony and Bartlett, 1999), and leads to bounds that have logarithmic dependencies on dimension. However, by observing that the effective dimension can be bounded by the sample size (as in Corollary 3), it is possible to remove this dimensional dependency for certain regularization conditions. This observation naturally leads to a new approach of using online learning to derive covering number bounds, as outlined in Section 3. Compared with earlier methods that have relied on the concept of "fat-shattering" dimension, our approach directly yields ∞-norm covering number bounds that improve previous results by a $\log n$ factor. Some specific consequences are discussed in Section 4.

Furthermore, the convex duality technique used in deriving online mistake bound (see Grove et al., 2001) is very general. It can be used to study general convex regularization

conditions. As an example, we are able to derive entropy covering number bounds that are difficult to obtained using previous techniques.

Also we shall mention that related to these covering number results, there have been some recent studies on randomized algorithms that select posterior distributions under certain regularization conditions (McAllester, 1999, Zhang, 1999). The generalization performance of these methods can be independent of the underlying dimension. Since these algorithms can be considered as special linear models, the dimension independent covering number bounds in Section 3 give intuitive explanations for the generalization ability of those algorithms within the traditional PAC analysis framework.

Acknowledgement

The author would like to thank Peter Bartlett and anonymous referees for pointing out related works, and for constructive suggestions that helped to improve the paper.

Appendix A. Proof of Theorem 2

For simplicity, we assume all quantities appearing in the proof are measurable. We follow the standard techniques (Pollard, 1984, Vapnik and Chervonenkis, 1971).

Step 1 (symmetrization by a replicate sample). For all $n\epsilon^2 \geq 2$, and consider i.i.d. random sample Y_1^n, independent of X_1^n,

$$\begin{aligned} &P\left[\sup_\alpha [E_{\mathbf{X}} f_1(\mathcal{L}(\alpha,\mathbf{x})) - E_{Y_1^n} f_2(\mathcal{L}(\alpha,\mathbf{y}))] > \epsilon\right] \\ \leq &2P\left[\sup_\alpha [E_{X_1^n} f_1(\mathcal{L}(\alpha,\mathbf{x})) - E_{Y_1^n} f_2(\mathcal{L}(\alpha,\mathbf{y}))] > \epsilon/2\right]. \end{aligned}$$

To see this, consider a function α^* such that $\alpha^*(Y_1^n)$ is a parameter that satisfies $E_{\mathbf{X}} f_1(\mathcal{L}(\alpha^*,\mathbf{x})) - E_{Y_1^n} f_2(\mathcal{L}(\alpha^*,\mathbf{y})) > \epsilon$ if such a parameter exists; and let $\alpha^*(Y_1^n)$ be an arbitrary parameter if no such parameter exists. Note that for any Y_1^n, by the Chebyshev's inequality, the conditional probability

$$\begin{aligned} &P\left[E_{\mathbf{X}} f_1(\mathcal{L}(\alpha^*,\mathbf{x})) - E_{X_1^n} f_1(\mathcal{L}(\alpha^*,\mathbf{x})) \leq \epsilon/2 | Y_1^n\right] \\ \geq &1 - \frac{1}{n\epsilon^2/4} E_{\mathbf{X}} f_1(\mathcal{L}(\alpha^*,\mathbf{x}))(1 - E_{\mathbf{X}} f_1(\mathcal{L}(\alpha^*,\mathbf{x}))) \geq 1/2. \end{aligned} \tag{5}$$

We thus have

$$\begin{aligned} &\frac{1}{2} P\left[\sup_\alpha [E_{\mathbf{X}} f_1(\mathcal{L}(\alpha,\mathbf{x})) - E_{Y_1^n} f_2(\mathcal{L}(\alpha,\mathbf{y}))] > \epsilon\right] \\ = &\frac{1}{2} P\left[E_{\mathbf{X}} f_1(\mathcal{L}(\alpha^*,\mathbf{x})) - E_{Y_1^n} f_2(\mathcal{L}(\alpha^*,\mathbf{y})) > \epsilon\right] \\ \leq &P\left[E_{\mathbf{X}} f_1(\mathcal{L}(\alpha^*,\mathbf{x})) - E_{Y_1^n} f_2(\mathcal{L}(\alpha^*,\mathbf{y})) > \epsilon, E_{\mathbf{X}} f_1(\mathcal{L}(\alpha^*,\mathbf{x})) - E_{X_1^n} f_1(\mathcal{L}(\alpha^*,\mathbf{y})) \leq \epsilon/2\right] \\ \leq &P\left[E_{X_1^n} f_1(\mathcal{L}(\alpha^*,\mathbf{x})) - E_{Y_1^n} f_2(\mathcal{L}(\alpha^*,\mathbf{y})) > \epsilon/2\right] \\ \leq &P\left[\sup_\alpha E_{X_1^n} f_1(\mathcal{L}(\alpha,\mathbf{x})) - E_{Y_1^n} f_2(\mathcal{L}(\alpha,\mathbf{y})) > \epsilon/2\right]. \end{aligned}$$

In the above derivation, the first inequality is a direct consequence of (5). The second and the third inequalities follow from simple algebra.

Step 2 (symmetrization by random signs). Consider i.i.d. sign variables $\sigma_1, \ldots, \sigma_n$, independent of X_1^n and Y_1^n, with $P(\sigma_i = -1) = P(\sigma_i = 1) = 1/2$. Define

$$g_\sigma(\alpha, \mathbf{x}) = (f_1(\mathcal{L}(\alpha, \mathbf{x})) - f_2(\mathcal{L}(\alpha, \mathbf{x})))/2 + \sigma(f_1(\mathcal{L}(\alpha, \mathbf{x})) + f_2(\mathcal{L}(\alpha, \mathbf{x})))/2,$$

and

$$h_\sigma(\alpha, \mathbf{y}) = -(f_1(\mathcal{L}(\alpha, \mathbf{y})) - f_2(\mathcal{L}(\alpha, \mathbf{y})))/2 + \sigma(f_1(\mathcal{L}(\alpha, \mathbf{y})) + f_2(\mathcal{L}(\alpha, \mathbf{y})))/2.$$

It is easy to check that

$$\begin{aligned}
&\sum_{i=1}^n [g_{\sigma_i}(\alpha, \mathbf{x}_i) - h_{\sigma_i}(\alpha, \mathbf{y}_i)] \\
=&\sum_{\sigma_i=1} [f_1(\mathcal{L}(\alpha, \mathbf{x}_i)) - f_2(\mathcal{L}(\alpha, \mathbf{y}_i))] + \sum_{\sigma_i=-1} [f_1(\mathcal{L}(\alpha, \mathbf{y}_i)) - f_2(\mathcal{L}(\alpha, \mathbf{x}_i))] .
\end{aligned}$$

This implies that the distribution of

$$\sup_\alpha \sum_{i=1}^n [f_1(\mathcal{L}(\alpha, \mathbf{x}_i)) - f_2(\mathcal{L}(\alpha, \mathbf{y}_i))]$$

is the same as that of

$$\sup_\alpha \sum_{i=1}^n [g_{\sigma_i}(\alpha, \mathbf{x}_i) - h_{\sigma_i}(\alpha, \mathbf{y}_i)] .$$

Therefore

$$\begin{aligned}
&P\left[\sup_\alpha E_{X_1^n} f_1(\mathcal{L}(\alpha, \mathbf{x})) - E_{Y_1^n} f_2(\mathcal{L}(\alpha, \mathbf{y})) > \epsilon/2\right] \\
=&P\left[\sup_\alpha \frac{1}{n} \sum_{i=1}^n (g_{\sigma_i}(\alpha, \mathbf{x}_i) - h_{\sigma_i}(\alpha, \mathbf{y}_i)) > \epsilon/2\right] \\
\leq&2P\left[\sup_\alpha \frac{1}{n} \sum_{i=1}^n g_{\sigma_i}(\alpha, \mathbf{x}_i) > \epsilon/4\right] .
\end{aligned}$$

Step 3 (derandomizing data). To estimate $P[\sup_\alpha \frac{1}{n} \sum_{i=1}^n g_{\sigma_i}(\alpha, \mathbf{x}_i) > \epsilon/4]$, we fix X_1^n and estimate the conditional probability

$$P\left[\sup_\alpha \frac{1}{n} \sum_{i=1}^n g_{\sigma_i}(\alpha, \mathbf{x}_i) > \epsilon/4 | X_1^n\right] .$$

Let $\{(\mathbf{z}_1^j, \ldots \mathbf{z}_n^j) : j = 1, \ldots, m\}$ be an infinity-norm γ-covering of $\mathcal{L}(\alpha, X_1^n)$, where $m = \mathcal{N}_\infty(\mathcal{L}, \gamma, X_1^n)$. By definition, for all α, there exists j such that $|\mathbf{z}_i^j - \mathcal{L}(\alpha, \mathbf{x}_i)| < \gamma$ for all i.

Therefore $g_1(\alpha, \mathbf{x}_i) = f_1(\mathcal{L}(\alpha, \mathbf{x}_i)) \leq f_3(\mathbf{z}_i^j)$ and $g_{-1}(\alpha, \mathbf{x}_i) = -f_2(\mathcal{L}(\alpha, \mathbf{x}_i)) \leq -f_3(\mathbf{z}_i^j)$; that is, $g_{\sigma_i}(\alpha, \mathbf{x}_i) \leq \sigma_i f_3(\mathbf{z}_i^j)$. We thus obtain

$$\begin{aligned}
&P\left[\sup_\alpha \frac{1}{n}\sum_{i=1}^n g_{\sigma_i}(\alpha, \mathbf{x}_i) > \epsilon/4 | X_1^n\right] \\
\leq &P\left[\sup_j \frac{1}{n}\sum_{i=1}^n \sigma_i f_3(\mathbf{z}_i^j) > \epsilon/4 | X_1^n\right] \\
\leq &\mathcal{N}_\infty(\mathcal{L}, \gamma, X_1^n) \sup_j P\left[\frac{1}{n}\sum_{i=1}^n \sigma_i f_3(\mathbf{z}_i^j) > \epsilon/4 | X_1^n\right] \\
\leq &\mathcal{N}_\infty(\mathcal{L}, \gamma, X_1^n) e^{-n\epsilon^2/32}.
\end{aligned}$$

The last inequality follows from the Hoeffding's inequality (Hoeffding, 1963). This proves the theorem.

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

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

A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Peter Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods : Support Vector Learning*, pages 43–54. The MIT press, 1999.

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(2):525–536, 1998.

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

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer-Verlag, New York, 1996. ISBN 0-387-94618-7.

R.M. Dudley. *A course on empirical processes*, volume 1097 of *Lecture Notes in Mathematics*. 1984.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

C. Gentile and M. K. Warmuth. Linear hinge loss and average margin. In *Proc. NIPS'98*, 1998.

A.J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43:173–210, 2001.

Ying Guo, Peter L. Bartlett, John Shawe-Taylor, and Robert C. Williamson. Covering numbers for support vector machines. In *COLT'99*, pages 267–277, 1999.

L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in banach spaces. In *Proceedings of Algorithmic Learning Theory*, pages 352–363, 1997.

D. Haussler. Generalizing the PAC model: sample size bounds from metric dimension-based uniform convergence results. In *Proc. 30th IEEE Symposium on Foundations of Computer Science*, pages 40–45, 1989.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 470–476. MIT Press, 2000.

Lee K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.*, 20(1): 608–613, 1992. ISSN 0090-5364.

J. Kivinen and M.K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Journal of Information and Computation*, 132:1–64, 1997.

A.N. Kolmogorov. Asymptotic characteristics of some completely bounded metric spaces. *Dokl. Akad. Nauk. SSSR*, 108:585–589, 1956.

A.N. Kolmogorov and V.M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.*, 17(2):277–364, 1961.

J. Langford and M. Seeger. Bounds for averaging classifiers. Technical Report CMU-CS-01-102, Carnegie Mellon University, 2001.

Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Springer-Verlag, Berlin, 1991. ISBN 3-540-52013-9. Isoperimetry and processes.

Wee Sun Lee, P.L. Bartlett, and R.C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

N. Littlestone. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

David McAllester. PAC-Bayesian model averaging. In *COLT'99*, pages 164–170, 1999.

Shahar Mendelson. Geometric methods in the analysis of Glivenko-Cantelli classes. In *COLT 01*, pages 256–272, 2001.

D. Pollard. *Convergence of stochastic processes*. Springer-Verlag, New York, 1984. ISBN 0-387-90990-7.

L.S. Pontriagin and L.G. Schnirelmann. Sur une propriété métrique de la dimension. *Annals of Mathematics*, 33:156–162, 1932.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (Series A)*, 13:145–147, 1972.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, 26(5):1651–1686, 1998. ISSN 0090-5364.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory*, 44(5):1926–1940, 1998.

Michael J. Todd. Potential-reduction methods in mathematical programming. *Math. Programming*, 76(1, Ser. B):3–45, 1997. ISSN 0025-5610. Interior point methods in theory and practice (Iowa City, IA, 1994).

V.N. Vapnik. *Statistical learning theory*. John Wiley & Sons, New York, 1998.

V.N. Vapnik and A.J. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Applications*, 16:264–280, 1971.

Robert C. Williamson, Alexander J. Smola, and Bernhard Schölkopf. Entropy numbers, operators and support vector kernels. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods : Support Vector Learning*, chapter 9. The MIT press, 1999.

Robert C. Williamson, Alexander J. Smola, and Bernhard Schölkopf. Entropy numbers of linear function classes. In *COLT'00*, pages 309–319, 2000.

Tong Zhang. Theoretical analysis of a class of randomized regularization methods. In *COLT 99*, pages 156–163, 1999.

Tong Zhang. On the dual formulation of regularized linear systems. *Machine Learning*, 46: 91–129, 2002.