Relative Deviation Margin Bounds

Corinna Cortes [1] **Mehryar Mohri** [1 2] **Ananda Theertha Suresh** [1]

Abstract

We present a series of new and more favorable margin-based learning guarantees that depend on the empirical margin loss of a predictor. We give two types of learning bounds, in terms of either the Rademacher complexity or the empirical ℓ_∞-covering number of the hypothesis set used, both distribution-dependent and valid for general families. Furthermore, using our relative deviation margin bounds, we derive distribution-dependent generalization bounds for unbounded loss functions under the assumption of a finite moment. We also briefly highlight several applications of these bounds and discuss their connection with existing results.

1. Introduction

Margin-based learning bounds provide a fundamental tool for the analysis of generalization in classification (Vapnik, 1998; 2006; Schapire et al., 1997; Koltchinskii and Panchenko, 2002; Taskar et al., 2003; Bartlett and Shawe-Taylor, 1998; Cortes et al., 2014; Kuznetsov et al., 2014; Cortes et al., 2017). These are guarantees that hold for real-valued functions based on the notion of confidence margin. Unlike worst-case bounds based on standard complexity measures such as the VC-dimension, margin bounds provide optimistic guarantees: a strong guarantee holds for predictors that achieve a relatively small empirical margin loss, for a relatively large value of the confidence margin. More generally, guarantees similar to margin bounds can be derived based on notion of a luckiness (Shawe-Taylor et al., 1998; Koltchinskii and Panchenko, 2002).

Notably, margin bounds do not have an explicit dependency on the dimension of the feature space for linear or kernel-based hypotheses. They provide strong guarantees for large-margin maximization algorithms such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995), including when they are used with positive definite kernels such as Gaussian kernels, for which the dimension of the feature space is infinite. Similarly, margin-based learning bounds have helped derive significant guarantees for AdaBoost (Freund and Schapire, 1997; Schapire et al., 1997). More recently, margin-based learning bounds have been derived for feed-forward artificial neural networks (NNs) (Neyshabur et al., 2015; Bartlett et al., 2017) and convolutional neural networks (CNNs) (Long and Sedghi, 2020).

An alternative family of tighter learning guarantees is that of relative deviation bounds (Vapnik, 1998; 2006; Anthony and Shawe-Taylor, 1993; Cortes et al., 2019). These are bounds on the difference of the generalization and the empirical error scaled by the square-root of the generalization error or empirical error, or some other power of the error. The scaling is similar to dividing by the standard deviation since, for smaller values of the error, the variance of the error of a predictor roughly coincides with its error. These guarantees translate into very useful bounds on the difference of the generalization error and empirical error whose complexity terms admit the empirical error as a factor.

This paper presents *relative deviation margin bounds*. These are new learning bounds that combine the benefit of standard margin bounds and that of standard relative deviation bounds, thereby resulting in tighter margin-based guarantees (Section 5.2). Our bounds are distribution-dependent and valid for general hypothesis sets. They can be viewed as "second-order" margin-based guarantees. For a sample size m, they are based on an interpolation between a $\frac{1}{\sqrt{m}}$-term that includes the square-root of the empirical margin loss as a factor and another term in $\frac{1}{m}$. In particular, when the empirical margin loss is zero, the bound only admits the $\frac{1}{m}$ fast rate term.

As an example, our learning bounds provide tighter guarantees for margin-based algorithms such as SVM and boosting than existing ones. We give two new families of relative deviation bounds, both distribution-dependent and valid for general hypothesis sets. Additionally, both families of guarantees hold for an arbitrary α-moment, with $\alpha \in (1, 2]$. The guarantees for general α-moments admit interesting applications in other areas. We describe one such application to deriving generalization guarantees for unbounded loss functions in Section 5.1.

[1]Google Research, New York, NY; [2]Courant Institute of Mathematical Sciences, New York, NY;. Correspondence to: Ananda Theertha Suresh <theertha@google.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Our first family of margin bounds are expressed in terms of the empirical ℓ_∞-covering number of the hypothesis set (Section 3). We show how these empirical covering numbers can be upper-bounded to derive empirical fat-shattering guarantees. One benefit of these resulting guarantees is that there are known upper bounds on the covering numbers for various standard hypothesis sets, which can be leveraged to derive explicit bounds.

Our second family of margin bounds are expressed in terms of the Rademacher complexity of the hypothesis set used (Section 4). Here, our learning bounds are first expressed in terms of a *peeling-based Rademacher complexity* term we introduce. Next, we give a series of upper bounds on this complexity measure, first simpler ones in terms of Rademacher complexity, next in terms of empirical ℓ_2-covering numbers, and finally in terms of the so-called *maximum Rademacher complexity*. In particular, we show that a simplified version of our bounds yields a guarantee similar to the maximum Rademacher margin bound of Srebro et al. (2010), but for a general α-moment.

We then use our families of margin bounds for α-moments to provide generalization guarantees for unbounded loss functions (Section 5.1). We also illustrate these results by deriving explicit bounds for various standard hypothesis sets in Section 5.2.

1.1. Contributions and Previous Work

We now further highlight our contributions and compare them to related previous work.

ℓ_∞-covering based bounds: A version of our main result for empirical ℓ_∞-covering number bounds in the special case $\alpha = 2$ was postulated by Bartlett (1998) without a proof. The author suggested that the proof could be given by combining various techniques with the results of Anthony and Shawe-Taylor (1993) and Vapnik (1998; 2006). However, as pointed out by Cortes et al. (2019), the proofs given by Anthony and Shawe-Taylor (1993) and Vapnik (1998; 2006) are incomplete and rely on a key lemma that is not proven by these authors. In a distinct line of research, Zhang (2002) presented finer covering number-based bounds for linear classifiers. These are not relative deviation bounds but the author postulated that his techniques could be modified, using Bernstein-type concentration bounds, to obtain relative deviation ℓ_∞-covering number bounds for linear classifiers. However, a careful inspection suggests that this is not a straightforward exercise and obtaining such bounds in fact requires techniques such as those we develop in this paper, or, perhaps, somewhat similar ones. *Our contribution:* We provide a self-contained proof based on a margin-based symmetrization argument. Our proof technique uses a new symmetrization argument that is different from those of Bartlett (1998) and Zhang (2002).

Rademacher complexity bounds: Using ideas from local Rademacher complexity (Bartlett et al., 2005), Rademacher complexity bounds were given by Srebro et al. (2010). However their bounds are based on the so-called *maximum Rademacher complexity*, which depends on the worst possible sample and is therefore independent of the underlying distribution. *Our contribution:* We provide the first distribution-dependent relative deviation margin bounds, in terms of a peeling-based Rademacher complexity. The proof is based on several new ingredients, including a new symmetrization result, an upper bound in terms of a normalized Rademacher process, and a peeling-based argument. We also show that, as a by-product of our guarantees, the distribution-independent bounds of Srebro et al. (2010) can be recovered, albeit with a more general $\alpha \in (1, 2]$.

Generalization bounds for unbounded loss functions: Standard relative deviation bounds do not hold for commonly used loss functions that are unbounded, such as cross-entropy. Cortes et al. (2019) provided zero-one relative deviation bounds which they used to derive guarantees for unbounded losses, in terms of the discrete dichotomies generated by the hypothesis class, under the assumption of a finite moment of the loss. *Our contribution*: We present the first generalization bounds for unbounded loss functions in terms of covering numbers and Rademacher complexity, which are *optimistic bounds* that, in general, are more favorable than the previous known bounds of Cortes et al. (2019), under the same finite moment assumption. Doing so further required us to derive relative deviation margin bounds for a general α-moment ($\alpha \in (1, 2]$), in contrast with previous work, which only focused on the special case $\alpha = 2$. The need for guarantees for unbounded loss functions with bounded α-moments with $\alpha < 2$ comes up in several scenarios, for example in the context of importance-weighting (Cortes, Mansour, and Mohri, 2010).

Recently, relative deviation margin bounds for the special case of linear classifiers were studied by Grønlund et al. (2020). Both the results and the proof techniques in that work are specific to the case of linear hypotheses. In contrast, our bounds hold for any general hypothesis set and recover the bounds of Grønlund et al. (2020) for the special case of linear classifiers, up to logarithmic factors. Furthermore, our proofs, while more general, are also simpler. Moreover, in contrast with these bounds, our guarantees are expressed in terms of Rademacher complexity and are therefore *distribution-dependent*. Relative deviation PAC-Bayesian bounds were also derived by McAllester (2003) for linear hypothesis sets. It is known, however, that Rademacher complexity learning bounds are finer guarantees: as shown by Kakade et al. (2008) and Foster et al. (2019)[Appendix H], they can be used to derive more favorable PAC-Bayesian guarantees than previously known ones (McAllester, 2003).

2. Preliminaries

In this section, we introduce the main definitions and notation used in our analysis and prove two symmetrization-type lemmas for a relative deviation between the expected binary loss and empirical margin loss.

We consider an input space $\mathcal{X}$ and a binary output space $\mathcal{Y} = \{-1, +1\}$ and a hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. We denote by $\mathcal{D}$ a distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and denote by $R(h)$ the generalization error and by $\widehat{R}_S(h)$ the empirical error of a hypothesis $h \in \mathcal{H}$:

$$R(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim\mathcal{D}}[1_{yh(x)\leq 0}],$$
$$\widehat{R}_S(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim S}[1_{yh(x)\leq 0}],$$

where we write $z \sim S$ to indicate that z is randomly drawn from the empirical distribution defined by S. Given $\rho \geq 0$, we similarly defined the ρ-margin loss and empirical ρ-margin loss of $h \in \mathcal{H}$:

$$R^\rho(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim\mathcal{D}}[1_{yh(x)<\rho}],$$
$$\widehat{R}^\rho_S(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim S}[1_{yh(x)<\rho}].$$

We will sometimes use the shorthand x_1^m to denote a sample of m points $(x_1, \ldots, x_m) \in \mathcal{X}^m$.

The *relative margin deviation* for a hypothesis $h \in \mathcal{H}$ is the ratio of the difference between the generalization error of h and its empirical margin loss, and the α-moment of the generalization error, $1 < \alpha \leq 2$:

$$\frac{R(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{R(h) + \tau}},$$

modulo a constant term $\tau > 0$ used to guarantee the positivity of denominator, which can be chosen to be arbitrarily small. For $R(h)$ small, the variance $R(h)(1 - R(h))$ is close to $R(h)$. Thus, for $\alpha = 2$, the ratio can be viewed as a normalization of the difference between the generalization error of h and its empirical margin loss obtained by dividing (approximately) by the standard deviation.

The problem we consider is to derive high-probability upper bounds for the supremum over $h \in \mathcal{H}$ of the relative margin deviation of h. This will result in our relative deviation margin bounds. We will be mainly interested in the case $\alpha = 2$. But, as we shall see in Section 5.1, the case $\alpha \in (1, 2)$ is crucial since it allows us to derive new covering number-based learning guarantees for unbounded loss functions when the α-moment of the loss is bounded only for some value $\alpha \in (1, 2)$.

The following is our first symmetrization lemma in terms of empirical margin losses. As already mentioned, the parameter $\tau > 0$ is used to ensure a positive denominator so that the relative deviations are mathematically well defined.

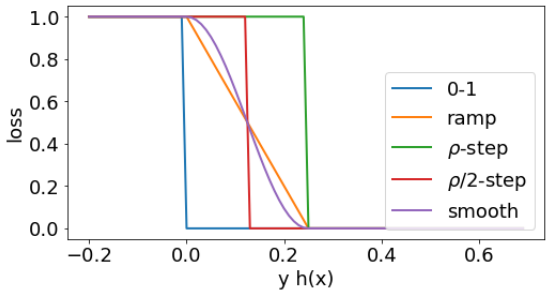

Figure 1. Illustration of different choices of function ϕ for $\rho = 0.25$.

Lemma 1. *Fix $\rho \geq 0$ and $1 < \alpha \leq 2$ and assume that $m\epsilon^{\frac{\alpha}{\alpha-1}} > 1$. Then, for any $\epsilon, \tau > 0$, the following inequality holds:*

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{R(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{R(h) + \tau}} > \epsilon\right]$$
$$\leq 4 \mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{\widehat{R}_{S'}(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(h) + \widehat{R}^\rho_S(h) + \frac{1}{m}]}} > \epsilon\right].$$

The proof is presented in Appendix A. It consists of extending the proof technique of Cortes et al. (2019) for standard empirical error to the empirical margin case and of using the binomial inequality (Greenberg and Mohri, 2013, Lemma 9). The lemma helps us bound the relative deviation in terms of the empirical margin loss on a sample S and the empirical error on an independent sample S', both of size m.

We now introduce some notation needed for the presentation and discussion of our relative deviation margin bound. Let $\phi: \mathbb{R} \to \mathbb{R}_+$ be a function such that the following inequality holds for all $x \in \mathbb{R}$:

$$1_{x<0} \leq \phi(x) \leq 1_{x<\rho}.$$

As an example, we can choose $\phi(x) = 1_{x<\rho/2}$ as in the previous sections. For a sample $z = (x, y)$, let $g(z) = \phi(yh(x))$. Then,

$$1_{yh(x)<0} \leq g(z) \leq 1_{yh(x)<\rho}. \tag{1}$$

Let the family $\mathcal{G}$ be defined as follows: $\mathcal{G} = \{z = (x, y) \mapsto \phi(yh(x)) : h \in \mathcal{H}\}$ and let $R(g) = \mathbb{E}_{z\sim\mathcal{D}}[g(z)]$ denote the expectation of g and $\widehat{R}_S(g) = \mathbb{E}_{z\sim S}[g(z)]$ its empirical expectation for a sample S. There are several choices for function ϕ, as illustrated by Figure 1. For example, $\phi(x)$ can be chosen to be $1_{x<\rho}$ or $1_{x<\rho/2}$ (Bartlett, 1998). ϕ can also be chosen to be the so-called *ramp loss*:

$$\phi(x) = \begin{cases} 1 & \text{if } x < 0 \\ 1 - \frac{x}{\rho} & \text{if } x \in [0, \rho] \\ 0 & \text{if } x > \rho, \end{cases}$$

or the smoothed margin loss chosen by (Srebro et al., 2010):

$$\phi(x) = \begin{cases} 1 & \text{if } x < 0 \\ \frac{1+\cos(\pi x/\rho)}{2} & \text{if } x \in [0, \rho] \\ 0 & \text{if } x > \rho. \end{cases}$$

Fix $\rho > 0$. Define the ρ-truncation function $\beta_\rho \colon \mathbb{R} \to [-\rho, +\rho]$ by $\beta_\rho(u) = \max\{u, -\rho\}1_{u \le 0} + \min\{u, +\rho\}1_{u \ge 0}$, for all $u \in \mathbb{R}$. For any $h \in \mathcal{H}$, we denote by h_ρ the ρ-truncation of h, $h_\rho = \beta_\rho(h)$, and define $\mathcal{H}_\rho = \{h_\rho \colon h \in \mathcal{H}\}$.

For any family of functions $\mathcal{F}$, we also denote by $\mathcal{N}_\infty(\mathcal{F}, \epsilon, x_1^m)$ the empirical covering number of $\mathcal{F}$ over the sample $(x_1, \ldots, x_m)$ and by $\mathcal{C}(\mathcal{F}, \epsilon, x_1^m)$ a minimum empirical cover. Then, the following symmetrization lemma holds.

Lemma 2. *Fix $\rho \ge 0$ and $1 < \alpha \le 2$. Then, the following inequality holds:*

$$\mathop{\mathbb{P}}_{S, S' \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} \frac{\widehat{R}_{S'}(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(h) + \widehat{R}^\rho_S(h) + \frac{1}{m}]}} > \epsilon\right] \le \mathop{\mathbb{P}}_{S, S' \sim \mathcal{D}^m}\left[\sup_{g \in \mathcal{G}} \frac{\widehat{R}_{S'}(g) - \widehat{R}_S(g)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(g) + \widehat{R}_S(g) + \frac{1}{m}]}} > \epsilon\right].$$

Further for $g(z) = 1_{yh(x) < \rho/2}$, using the shorthand $\mathcal{K} = \mathcal{C}(\mathcal{H}_\rho, \frac{\rho}{2}, S \cup S')$, the following holds:

$$\mathop{\mathbb{P}}_{S, S' \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} \frac{\widehat{R}_{S'}(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(h) + \widehat{R}^\rho_S(h) + \frac{1}{m}]}} > \epsilon\right] \le \mathop{\mathbb{P}}_{S, S' \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{K}} \frac{\widehat{R}^{\frac{\rho}{2}}_{S'}(h) - \widehat{R}^{\frac{\rho}{2}}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}^{\frac{\rho}{2}}_{S'}(h) + \widehat{R}^{\frac{\rho}{2}}_S(h) + \frac{1}{m}]}} > \epsilon\right].$$

The proof consists of using Inequality 1, it is given in Appendix A. The first result of the lemma gives an upper bound for a general choice of functions g, that is for an arbitrary choices of the Φ loss function. This inequality will be used in Section 4 to derive our Rademacher complexity bounds. The second inequality is for the specific choice of Φ that corresponds to $\rho/2$-step function. We will use this inequality in the next section to derive ℓ_∞-covering number bounds.

3. Relative Deviation Margin Bounds – Covering Numbers

In this section, we present a general relative deviation margin-based learning bound, expressed in terms of the expected empirical covering number of $\mathcal{H}_\rho$. The learning guarantee is thus distribution-dependent. It is also very general since it is given for any $1 < \alpha \le 2$ and an arbitrary hypothesis set.

Theorem 1 (General relative deviation margin bound)**.** *Fix $\rho \ge 0$ and $1 < \alpha \le 2$. Then, for any hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and any $\tau > 0$, the following inequality holds:*

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} \frac{R(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{R(h) + \tau}} > \epsilon\right] \le 4 \mathop{\mathbb{E}}_{x_1^{2m} \sim \mathcal{D}^{2m}}[\mathcal{N}_\infty(\mathcal{H}_\rho, \tfrac{\rho}{2}, x_1^{2m})] \exp\left[\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right].$$

The proof is given in Appendix B. As mentioned earlier, a version of this result for $\alpha = 2$ was postulated by Bartlett (1998). The result can be alternatively expressed as follows, taking the limit $\tau \to 0$.

Corollary 1. *Fix $\rho \ge 0$ and $1 < \alpha \le 2$. Then, for any hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, with probability at least $1 - \delta$, the following inequality holds for all $h \in \mathcal{H}$:*

$$R(h) - \widehat{R}^\rho_S(h) \le 2^{\frac{\alpha+2}{2\alpha}} \sqrt[\alpha]{R(h)} \sqrt{\frac{\log \mathbb{E}[\mathcal{N}_\infty(\mathcal{H}_\rho, \frac{\rho}{2}, x_1^{2m})] + \log \frac{1}{\delta}}{m^{\frac{2(\alpha-1)}{\alpha}}}}.$$

Note that a smaller value of α (α closer to 1) might be advantageous for some values of $R(h)$, at the price of a worse complexity in terms of the sample size. For $\alpha = 2$, the result can be rewritten as follows. In the following, we use $\overline{\mathcal{N}}_\infty$ as a shorthand for $\mathbb{E}[\mathcal{N}_\infty(\mathcal{H}_\rho, \frac{\rho}{4}, x_1^{2m})]$.

Corollary 2. *Fix $\rho \ge 0$. Then, for any hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, with probability at least $1 - \delta$, the following inequality holds for all $h \in \mathcal{H}$:*

$$R(h) - \widehat{R}^\rho_S(h) \le 2\sqrt{\widehat{R}^\rho_S(h)\frac{\log \frac{\overline{\mathcal{N}}_\infty}{\delta}}{m}} + 4\frac{\overline{\mathcal{N}}_\infty}{\delta}{m}.$$

Proof. Let a, b, and c be defined as follows: $a = R(h)$, $b = \widehat{R}^\rho_S(h)$, and $c = \frac{\log \mathbb{E}[\mathcal{N}_\infty(\mathcal{H}_\rho, \frac{\rho}{2}, x_1^{2m})), \frac{\rho}{2})] + \log \frac{1}{\delta}}{m}$. Then, for $\alpha = 2$, the inequality of Corollary 1 can be rewritten as

$$a \le b + 2\sqrt{ca}.$$

This implies that $(\sqrt{a} - \sqrt{c})^2 \le b + c$ and hence $\sqrt{a} \le \sqrt{b + c} + \sqrt{c}$. Therefore, $a \le b + 2c + 2\sqrt{(b + c)c} \le b + 4c + 2\sqrt{cb}$. Substituting the values of a, b, and c yields the bound. □

The guarantee just presented provides a tighter margin-based learning bound than standard margin bounds since the dominating term admits the empirical margin loss as a factor. Standard margin bounds are subject to a trade-off: a large

value of ρ reduces the complexity term while leading to a larger empirical margin loss term. Here, the presence of the empirical loss factor favors this trade-off by allowing a smaller choice of ρ. The bound is distribution-dependent since it is expressed in terms of the expected covering number and it holds for an arbitrary hypothesis set $\mathcal{H}$.

The learning bounds just presented hold for a fixed value of ρ. They can be extended to hold uniformly for all values of $\rho \in [0,1]$, at the price of an additional $\log\log$-term. We illustrate that extension for Corollary 1.

Corollary 3. *Fix $1 < \alpha \leq 2$. Then, for any hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and any $\rho \in (0, r]$, with probability $\geq 1-\delta$, the following inequality holds for all $h \in \mathcal{H}$:*

$$R(h) \leq \widehat{R}_S^{\rho}(h) + 2^{\frac{\alpha+2}{2\alpha}} \sqrt[\alpha]{R(h)} \sqrt{\frac{\log \overline{\mathcal{N}}_\infty + \log\left(\frac{\log_2(2r/\rho)}{\delta}\right)}{m^{\frac{2(\alpha-1)}{\alpha}}}}.$$

Proof. For $k \geq 1$, let $\rho_k = r/2^k$ and $\delta_k = \delta/k^2$. For all such ρ_k, by Corollary 1 and the union bound,

$$R(h) \leq \widehat{R}_S^{\rho_k}(h) + 2^{\frac{\alpha+2}{2\alpha}} \sqrt[\alpha]{R(h)} \sqrt{\frac{\log \overline{\mathcal{N}}_\infty + \log\frac{1}{\delta} + 2\log k}{m^{\frac{2(\alpha-1)}{\alpha}}}}.$$

By the union bound, the error probability is most $\sum_k \delta_k = \delta \sum_k (1/k^2) \leq \delta$. For any $\rho \in (0, r]$, there exists a k such that $\rho \in (\rho_k, \rho_{k-1}]$. For this k, $\rho \leq \rho_{k-1} = r/2^{k-1}$. Hence, $k \leq \log_2(2r/\rho)$. By the definition of margin, for all $h \in \mathcal{H}$, $\widehat{R}_S^{\rho_k}(h) \leq \widehat{R}_S^{\rho}(h)$. Furthermore, as $\rho_k = \rho_{k-1}/2 \geq \rho/2$, $\mathcal{N}_\infty(\mathcal{H}_\rho, \frac{\rho_k}{2}, x_1^{2m}) \leq \mathcal{N}_\infty(\mathcal{H}_\rho, \frac{\rho}{4}, x_1^{2m})$. Hence, for all $\rho \in (0, r]$,

$$R(h) \leq \widehat{R}_S^{\rho}(h) + 2^{\frac{\alpha+2}{2\alpha}} \sqrt[\alpha]{R(h)} \sqrt{\frac{\log \overline{\mathcal{N}}_\infty + \log\left(\frac{\log_2(2r/\rho)}{\delta}\right)}{m^{\frac{2(\alpha-1)}{\alpha}}}}.$$

This concludes the proof. □

Our previous bounds can be expressed in terms of the fat-shattering dimension, as illustrated below. Recall that, given $\gamma > 0$, a set of points $\mathcal{U} = \{u_1, \ldots, u_m\}$ is said to be γ-*shattered* by a family of real-valued functions $\mathcal{H}$ if there exist real numbers $(r_1, \ldots, r_m)$ (witnesses) such that for all binary vectors $(b_1, \ldots, b_m) \in \{0,1\}^m$, there exists $h \in \mathcal{H}$ such that:

$$h(x) \begin{cases} \geq r_i + \gamma & \text{if } b_i = 1; \\ \leq r_i - \gamma & \text{otherwise.} \end{cases}$$

The *fat-shattering dimension* $\mathrm{fat}_\gamma(\mathcal{H})$ of the family $\mathcal{H}$ is the cardinality of the largest set γ-shattered set by $\mathcal{H}$ (Anthony and Bartlett, 1999).

Corollary 4. *Fix $\rho \geq 0$. Then, for any hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ with $d = \mathrm{fat}_{\frac{\rho}{16}}(\mathcal{H})$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$:*

$$R(h) \leq \widehat{R}_S^{\rho}(h) + 2\sqrt{\widehat{R}_S^{\rho}(h)\frac{\Delta_m}{m}} + \frac{\Delta_m}{m},$$

where $\Delta_m = 1 + d\log_2(2c^2m)\log_2\frac{2cem}{d} + \log\frac{1}{\delta}$ and $c = 17$.

Proof. By (Bartlett, 1998, Proof of theorem 2), we have

$$\log \max_{x_1^{2m}} [\mathcal{N}_\infty(\mathcal{H}_\rho, \tfrac{\rho}{2}, x_1^{2m})] \leq 1 + d' \log_2(2c^2m)\log_2\frac{2cem}{d'},$$

where $d' = \mathrm{fat}_{\frac{\rho}{16}}(\mathcal{H}_\rho) \leq \mathrm{fat}_{\frac{\rho}{16}}(\mathcal{H}) = d$. Upper bounding the expectation by the maximum completes the proof. □

We will use this bound in Section 5.2 to derive explicit guarantees for several standard hypothesis sets.

4. Relative Deviation Margin Bounds – Rademacher Complexity

In this section, we present relative deviation margin bounds expressed in terms of the Rademacher complexity of the hypothesis sets. As with the previous section, these bounds are general: they hold for any $1 < \alpha \leq 2$ and arbitrary hypothesis sets.

As in the previous section, we will define the family $\mathcal{G}$ by $\mathcal{G} = \{\phi(yh(x)) : h \in \mathcal{H}\}$, where ϕ is a function such that

$$1_{x<0} \leq \phi(x) \leq 1_{x<\rho}.$$

For a set $\mathcal{G}$ and a set of samples z_1^m, the empirical Rademacher complexity is defined as

$$\widehat{\mathfrak{R}}_m(\mathcal{G}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_i \sigma_i g(z_i)\right].$$

We further allow $\mathcal{G}$ to be dependent on the samples.

The proof of our main result in this section admits the following three main ingredients: (1) a symmetrization lemma to relate the relative margin deviation term to a symmetrized quantity with empirical terms only (Lemmas 1 and 2); (2) relating the problem of bounding that symmetrized quantity to that of bounding a normalized Rademacher process (Lemma 3); (3) bounding that normalized Rademacher process in terms of Rademacher complexity using an adapted peeling technique.

4.1. Rademacher Complexity-Based Margin Bounds

We first relate bounding the symmetrized relative deviations to bounding the *normalized Rademacher process* $\sup_{g\in\mathcal{G}} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m g(z_i)+1]}}$.

Lemma 3. *Fix $1 < \alpha \leq 2$. Then, the following inequality holds:*

$$\mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{g\in\mathcal{G}} \frac{\widehat{R}_{S'}(g) - \widehat{R}_S(g)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(g) + \widehat{R}_S(g) + \frac{1}{m}]}} > \epsilon\right] \leq 2 \mathop{\mathbb{P}}_{z_1^m\sim\mathcal{D}^m,\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m g(z_i) + 1]}} > \frac{\epsilon}{2\sqrt{2}}\right].$$

The proof is given in Appendix C. It consists of introducing Rademacher variables and deriving an upper bound in terms of the first m points only.

Now, to bound the normalized Rademacher process term, the technique adopted in previous work has consisted of fixing z_1^m and applying Hoeffding's bound to the ratio $\frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i))+1]}}$ for a fixed $g \in \mathcal{G}$ (Anthony and Shawe-Taylor, 1993; Cortes et al., 2019). This is then followed by a union bound, which results in shattering coefficients or covering numbers, and an expectation over z_1^m.

Instead, for a fixed z_1^m, we will seek to directly bound the normalized Rademacher process term via a uniform convergence bound. Doing so is not straightforward due to the complex denominator. Thus, we first *peel* $\mathcal{G}$ according to the values of the main term in the denominator $\frac{1}{m}\sum_{i=1}^m g(z_i) + \frac{1}{m}$: we partition $\mathcal{G}$ into sets $\mathcal{G}_k(z_1^m)$ for which this average value is in $\left[\frac{2^k}{m}, \frac{2^{k+1}-1}{m}\right]$. This reduces bounding the normalized Rademacher process term to that of bounding the Rademacher process terms $\sup_{g\in\mathcal{G}_k(z_1^m)} \frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)$. Now, to bound these terms, using McDiarmid's inequality would result in too loose terms. This is essentially because the proxy term for the variance in McDiarmid's inequality is a quantity of the form $\sum_{i=1}^m \|\Delta_i g\|_\infty^2$. We use an alternative bounded difference inequality (van Handel, 2016, Theorem 3.18) with a proxy term of the form $\|\sum_{i=1}^m \Delta_i g\|_\infty^2$ instead, which helps us leverage the property of $\mathcal{G}_k(z_1^m)$ and also provide a finer one-sided inequality. This results, for each Rademacher process term $\sup_{g\in\mathcal{G}_k(z_1^m)} \frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)$, in a bound expressed in terms of the Rademacher complexity of $\mathcal{G}_k(z_1^m)$. A union bound over the sets $\mathcal{G}_k(z_1^m)$ and an expectation over z_1^m conclude the proof.

With this background, we now detail the peeling argument, that is we partition $\mathcal{G}$ into subsets $\mathcal{G}_k$, give a learning bound for each $\mathcal{G}_k$, and then take a weighted union bound. For any non-negative integer k with $0 \leq k \leq \log_2 m$, let $\mathcal{G}_k(z_1^m)$ denote the family of hypotheses defined by

$$\mathcal{G}_k(z_1^m) = \left\{g \in \mathcal{G} : 2^k \leq \left(\sum_{i=1}^m g(z_i)\right) + 1 < 2^{k+1}\right\}.$$

Using the above inequality and a peeling argument, we show the following upper bound expressed in terms of Rademacher complexities.

Lemma 4. *Fix $1 < \alpha \leq 2$ and $z_1^m \in \mathcal{Z}^m$. Then, the following inequality holds:*

$$\mathop{\mathbb{P}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i)) + 1]}} > \epsilon \,\middle|\, z^m\right] \leq 2\sum_{k=0}^{\lfloor \log_2 m\rfloor} \exp\left[\frac{m^2\widehat{\mathfrak{R}}_m^2(\mathcal{G}_k(z_1^m))}{2^{k+5}} - \frac{\epsilon^2}{64\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right] 1_{\epsilon\leq 2[\frac{2^k}{m}]^{1-\frac{1}{\alpha}}}.$$

The proof is given in Appendix C. Instead of applying Hoeffding's bound to each term of the left-hand side for a fixed g and then using covering and the union bound to bound the supremum, here, we seek to bound the supremum over $\mathcal{G}$ directly. To do so, we use a bounded difference inequality that leads to a finer result than McDiarmid's inequality.

Let $\mathfrak{r}_m(\mathcal{G})$ be defined as the following *peeling-based Rademacher* complexity of $\mathcal{G}$:

$$\sup_{0\leq k\leq \log_2(m)} \log\left[\mathop{\mathbb{E}}_{z_1^m\sim\mathcal{D}^m}\left[\exp\left(\frac{m^2\widehat{\mathfrak{R}}_m^2(\mathcal{G}_k(z_1^m))}{2^{k+5}}\right)\right]\right].$$

Then, the following is a margin-based relative deviation bound expressed in terms of $\mathfrak{r}_m(\mathcal{G})$, that is in terms of Rademacher complexities.

Theorem 2. *Fix $1 < \alpha \leq 2$. Then, with probability at least $1-\delta$, for all hypothesis $h \in \mathcal{H}$, the following inequality holds:*

$$R(h) - \widehat{R}_S^\rho(h) \leq 16\sqrt{2}\sqrt[\alpha]{R(h)}\left[\frac{\mathfrak{r}_m(\mathcal{G}) + \log\log m + \log\frac{16}{\delta}}{m}\right]^{1-\frac{1}{\alpha}}.$$

Combining the above lemma with Theorem 2 yields the following.

Corollary 5. *Fix $1 < \alpha \leq 2$ and let $\mathcal{G}$ be defined as above. Then, with probability at least $1-\delta$, for all hypothesis $h \in \mathcal{H}$,*

$$R(h) - \widehat{R}_S^\rho(h) \leq 32\sqrt[\alpha]{\widehat{R}_S^\rho(h)}\left(\frac{\Delta_m}{m}\right)^{1-\frac{1}{\alpha}} + 2(32)^{\frac{\alpha}{\alpha-1}}\left(\frac{\Delta_m}{m}\right),$$

where $\Delta_m = \mathfrak{r}_m(\mathcal{G}) + \log\log m + \log\frac{16}{\delta}$.

The above result can be extended to hold for all α simultaneously.

Corollary 6. *Let $\mathcal{G}$ be defined as above. Then, with probability at least $1-\delta$, for all hypothesis $h \in \mathcal{H}$ and $\alpha \in (1, 2]$,*

$$R(h) - \widehat{R}_S^\rho(h) \leq 32\sqrt{2}\sqrt[\alpha]{R(h)}\left[\frac{\mathfrak{r}_m(\mathcal{G}) + \log\frac{16\log m}{\delta}}{m}\right]^{1-\frac{1}{\alpha}}.$$

4.2. Upper Bounds on Peeling-Based Rademacher Complexity

We now present several upper bounds on $\mathfrak{r}_m(\mathcal{G})$ and show how this can help recover previously known quantities. We provide proofs for all the results in Appendix D. For any hypothesis set $\mathcal{G}$, we denote by $\mathbb{S}_{\mathcal{G}}(x_1^m)$ the number of distinct dichotomies generated by $\mathcal{G}$ over that sample:

$$\mathbb{S}_{\mathcal{G}}(z_1^m) = \mathrm{Card}\big(\big\{\big(g(z_1),\ldots,g(z_m)\big) : g \in \mathcal{G}\big\}\big).$$

We note that we do not make any assumptions over the range of $\mathcal{G}$.

Lemma 5. *If the functions in $\mathcal{G}$ take values in $\{0,1\}$, then the following upper bounds hold for the peeling-based Rademacher complexity of $\mathcal{G}$:*

$$\mathfrak{r}_m(\mathcal{G}) \le \frac{1}{8}\log \mathop{\mathbb{E}}_{z_1^m}[\mathbb{S}_{\mathcal{G}}(z_1^m)].$$

Combining the above result with Corollary 5, improves the relative deviation bounds of (Cortes et al., 2019, Corollary 2) for $\alpha < 2$. In particular, we improve the $\sqrt{\mathbb{E}_{z_1^m}[\mathbb{S}_{\mathcal{G}}(z_1^m)]}$ term in their bounds to $\big(\mathbb{E}_{z_1^m}[\mathbb{S}_{\mathcal{G}}(z_1^m)]\big)^{1-1/\alpha}$, which is significant for $\alpha < 2$.

We next upper bound the peeling-based Rademacher complexity in terms of covering numbers.

Lemma 6. *For a set of hypotheses $\mathcal{G}$,*

$$\mathfrak{r}_m(\mathcal{G}) \le \sup_{0\le k\le \log_2(m)} \log\left[\mathop{\mathbb{E}}_{z_1^m\sim\mathcal{D}^m}\big[\exp\big\{f_k(z_1^m,\mathcal{G})\big\}\big]\right].$$

where

$$f_k(z_1^m,\mathcal{G}) = \frac{1}{16}\left[1 + \int_{\frac{1}{\sqrt{m}}}^{1} \log \mathcal{N}_2\left(\mathcal{G}_k(z_1^m), \sqrt{\tfrac{2^k}{m}}\epsilon, z_1^m\right) d\epsilon\right].$$

One can further simplify the above bound using the smoothed margin loss from (Srebro et al., 2010). Let the worst case Rademacher complexity be defined as follows.

$$\widehat{\mathfrak{R}}_m^{\max}(\mathcal{H}) = \sup_{z_1^m} \widehat{\mathfrak{R}}_m(\mathcal{H}).$$

Lemma 7. *Let g be the smoothed margin loss from (Srebro et al., 2010, Section 5.1), with its second moment bounded by $(\pi^2/4\rho^2)$. Then, $\mathfrak{r}_m(\mathcal{G})$ is upper bounded by*

$$\frac{\big[4\pi\widehat{\mathfrak{R}}_m^{\max}(\mathcal{H})\big]^2}{(\rho^2/m)}\left[2\log^{\frac{3}{2}}\left[\frac{m}{\widehat{\mathfrak{R}}_m^{\max}(\mathcal{H})}\right] - \log^{\frac{3}{2}}\left[\frac{2\pi m}{\rho\widehat{\mathfrak{R}}_m^{\max}(\mathcal{H})}\right]\right]^2.$$

Proof. Recall that the smoothed margin loss of Srebro et al. (2010) is given by

$$g(yh(x)) = \begin{cases} 1 & \text{if } yh(x) < 0 \\ \frac{1+\cos(\pi yh(x)/\rho)}{2} & \text{if } yh(x) \in [0,\rho] \\ 0 & \text{if } yh(x) > \rho. \end{cases}$$

Upper bounding the expectation by the maximum gives:

$$\mathfrak{r}_m(\mathcal{G}) \le \sup_k \log \sup_{z_1^m}\left[\exp\left(\frac{m^2\widehat{\mathfrak{R}}_m^2(\mathcal{G}_k(z_1^m))}{2^{k+5}}\right)\right] \le \sup_k \sup_{z_1^m} \frac{m^2\widehat{\mathfrak{R}}_m^2(\mathcal{G}_k(z_1^m))}{2^{k+5}}.$$

Let $\mathcal{G}'_k(z_1^m) = \left\{g \in \mathcal{G} : \sum_{i=1}^m g(z_i) + 1 \le 2^{k+1}\right\}$. Since $\mathcal{G}_k(z_1^m) \subseteq \mathcal{G}'_k(z^m)$,

$$\mathfrak{r}_m(\mathcal{G}) \le \sup_k \sup_{z_1^m} \frac{m^2\widehat{\mathfrak{R}}_m^2(\mathcal{G}'_k(z^m))}{2^{k+5}}.$$

Now, $\widehat{\mathfrak{R}}_m(\mathcal{G}'_k(z^m))$ coincides with the local Rademacher complexity term defined in (Srebro et al., 2010, Section 2). Thus, by (Srebro et al., 2010, Lemma 2.2), $\frac{\widehat{\mathfrak{R}}_m(\mathcal{G}'_k(z^m))}{\widehat{\mathfrak{R}}_m^{\max}(\mathcal{H})}$ is upper bounded by

$$\frac{16\pi}{\rho}\sqrt{\frac{2^{k+1}}{m}}\left[2\log^{\frac{3}{2}}\left[\frac{m}{\widehat{\mathfrak{R}}_m^{\max}(\mathcal{H})}\right] - \log^{\frac{3}{2}}\left[\frac{2\pi m}{\rho\widehat{\mathfrak{R}}_m^{\max}(\mathcal{H})}\right]\right],$$

which concludes the proof. □

Combining Lemma 7 with Corollary 5 yields the following bound, which is a generalization of (Srebro et al., 2010, Theorem 5) holding for all $\alpha \in (1,2]$.

Corollary 7. *For any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds for all $\alpha \in (0,1]$ and all $h \in \mathcal{H}$:*

$$R(h) - \widehat{R}_S^{\rho}(h) \le 32\sqrt{2}\sqrt[\alpha]{\widehat{R}_S^{\rho}(h)}\,\beta_m^{1-\frac{1}{\alpha}} + 2(32)^{\frac{\alpha}{\alpha-1}}\beta_m,$$

where β_m is the upper bound on $\mathfrak{r}_m(\mathcal{G})$ in Lemma 7.

5. Applications

In this section, we discuss two applications of our relative deviation margin bounds. We first show how they can be used to obtain generalization guarantees for unbounded loss functions. Next, we describe the application of our bounds to several specific hypothesis sets and show they can recover some recent results. In Appendix F, we further discuss other potential applications of our learning guarantees.

5.1. Generalization Bounds for Unbounded Loss Functions

Standard generalization bounds hold for bounded loss functions. Many loss functions frequently used in applications, such as the cross-entropy loss, are unbounded, when used with standard hypothesis sets. For the more general and more realistic case of unbounded loss functions, a number of different results have been presented in the past, under different assumption on the family of functions. This includes

learning bounds assuming the existence of an *envelope*, that is a single non-negative function with a finite expectation lying above the absolute value of the loss of every function in the hypothesis set (Dudley, 1984; Pollard, 1984; Dudley, 1987; Pollard, 1989; Haussler, 1992), or an assumption similar to Hoeffding's inequality based on the expectation of a hyperbolic function, a quantity similar to the moment-generating function (Meir and Zhang, 2003), or the weaker assumption that the αth-moment of the loss is bounded for some value of $\alpha > 1$ (Vapnik, 1998; 2006; Cortes et al., 2019). The need for guarantees for unbounded loss functions with bounded alpha-moments with $\alpha < 2$ come up in several scenarios, for example in the context of importance-weighting (Cortes, Mansour, and Mohri, 2010). Here, we will also adopt this assumption and present distribution-dependent learning bounds for unbounded losses that improve upon the previous bounds of Cortes et al. (2019). To do so, we will leverage the relative deviation margin bounds given in the previous sections, which hold for any $\alpha \leq 2$.

Let L be an unbounded loss function and $L(h, z)$ denote the loss of hypothesis h for sample z. Let $\mathcal{L}_\alpha(h) = \mathbb{E}_{z\sim D}[L(h,z)^\alpha]$ be the α^{th}-moment of the loss function $\mathcal{L}$, which is assumed finite for all $h \in \mathcal{H}$. In what follows, we will use the shorthand $\mathbb{P}[L(h,z) > t]$ instead of $\mathbb{P}_{z\sim D}[L(h,z) > t]$, and similarly $\widehat{\mathbb{P}}[L(h,z) > t]$ instead of $\mathbb{P}_{z\sim \widehat{D}}[L(h,z) > t]$.

Theorem 3. *Fix $\rho \geq 0$. Let $1 < \alpha \leq 2$, $0 < \epsilon \leq 1$, and $0 < \tau^{\frac{\alpha-1}{\alpha}} < \epsilon^{\frac{\alpha}{\alpha-1}}$. For any loss function L (not necessarily bounded) and hypothesis set $\mathcal{H}$ such that $\mathcal{L}_\alpha(h) < +\infty$ for all $h \in \mathcal{H}$,*

$$\mathbb{P}\left[\sup_{h\in H} \mathcal{L}(h) - \widehat{\mathcal{L}}_S(h) > \Gamma_\tau(\alpha,\epsilon)\, \epsilon \sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau} + \rho\right] \leq \mathbb{P}\left[\sup_{h\in\mathcal{H}, t\in\mathbb{R}} \frac{\mathbb{P}[L(h,z) > t] - \widehat{\mathbb{P}}[L(h,z) > t - \rho]}{\sqrt[\alpha]{\mathbb{P}[L(h,z) > t] + \tau}} > \epsilon\right],$$

where $\Gamma_\tau(\alpha,\epsilon) = \frac{\alpha-1}{\alpha}(1+\tau)^{\frac{1}{\alpha}} + \frac{1}{\alpha}\left(\frac{\alpha}{\alpha-1}\right)^{\alpha-1}(1 + \left(\frac{\alpha-1}{\alpha}\right)^\alpha \tau^{\frac{1}{\alpha}})^{\frac{1}{\alpha}} \left[1 + \frac{\log(1/\epsilon)}{(\frac{\alpha}{\alpha-1})^{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$.

The proof is provided in Appendix E. The above theorem can be used in conjunction with our relative deviation margin bounds to obtain strong guarantees for unbounded loss functions and we illustrate it with our ℓ_∞-based bounds. Similar techniques can be used to obtain peeling-based Rademacher complexity bounds. Combining Theorems 3 and (1) yields the following corollary.

Corollary 8. *Fix $\rho \geq 0$. Let $\epsilon < 1$, $1 < \alpha \leq 2$. and hypothesis set $\mathcal{H}$ such that $\mathcal{L}_\alpha(h) < +\infty$ for all $h \in \mathcal{H}$,*

$$\mathcal{L}(h) - \widehat{\mathcal{L}}_S(h) \leq \gamma \sqrt[\alpha]{\mathcal{L}_\alpha(h)} \sqrt{\frac{\Delta_m}{m^{\frac{2(\alpha-1)}{\alpha}}}} + \rho,$$

where $\Delta_m = \log \mathbb{E}[\mathcal{N}_\infty(\mathcal{L}(\mathcal{H}), \frac{\rho}{2}, x_1^{2m})] + \log\frac{1}{\delta}$ *and* $\gamma = \Gamma_0\left(\alpha, \sqrt{\frac{\Delta_m}{m^{\frac{2(\alpha-1)}{\alpha}}}}\right) = \mathcal{O}(\log m)$.

The upper bound in the above corollary has two terms. The first term is based on the covering number and decreases with ρ while the second term increases with ρ. A natural choice for ρ is $1/\sqrt{m}$, however one can choose a suitable value of ρ that minimizes the sum to obtain favorable bounds.[1] Furthermore, the above bound depends on the covering number as opposed to the result of Cortes et al. (2019), which depends on the number of dichotomies generated by the hypothesis set. Hence, the above bound is *optimistic* and in general is more favorable than the previous known bounds of Cortes et al. (2019). We note that instead of using our ℓ_∞-based bounds, one can use our Rademacher complexity bounds to derive finer results.

5.2. Relative Margin Bounds for Common Hypothesis Sets

In this section, we briefly highlight some applications of our learning bounds: both our covering number and Rademacher complexity margin bounds can be used to derive finer margin-based guarantees for several commonly used hypothesis sets. Below we briefly illustrate these applications.

Linear hypothesis sets: let $\mathcal{H}$ be the family of liner hypotheses defined by

$$\mathcal{H} = \{\mathbf{x} \mapsto \mathbf{w}\cdot\mathbf{x} \colon \|\mathbf{w}\|_2 \leq 1, \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2 \leq R\}.$$

The margin bound for SVM by Bartlett and Shawe-Taylor (1998, Theorem 1.7) is

$$R(h) \leq \widehat{R}_S^\rho(h) + c'\sqrt{\beta'_m}, \tag{2}$$

where c' is some universal constant and where $\beta'_m = \widetilde{O}\left(\frac{(R/\rho)^2}{m}\right)$. Recently, Grønlund et al. (2020) derived the following more favorable relative deviation margin bounds for linear hypothesis sets:

$$R(h) \leq \widehat{R}_S^\rho(h) + 2\sqrt{\widehat{R}_S^\rho(h)\,\beta''_m} + \beta''_m, \tag{3}$$

where $\beta''_m = \widetilde{O}\left(\frac{(R/\rho)^2}{m}\right)$. We can directly apply our relative deviation margin bounds to recover this result up to logarithmic factors. However, our guarantees have the additional benefit of being expressed in terms of Rademacher complexity and thus of being distribution-dependent, unlike the bound of Grønlund et al. (2020). Furthermore, while their proof technique crucially depends on the fact that the underlying hypothesis set is linear, ours is comparatively

[1] This requires that the bound holds uniformly for all ρ, which can be shown with an additional $\log\log\frac{1}{\rho}$ term (See Corollary 9).

simpler and very general, it applies to arbitrary hypothesis sets.

Feed-forward neural networks of depth d: For a matrix $\mathbf{W}$, let $\|\mathbf{W}\|_{p,q}$ denote the matrix p,q norm and $\|\mathbf{W}\|_2$ denote the spectral norm. Let $\mathcal{H}_0 = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1, \mathbf{x} \in \mathbf{R}^n\}$ and $\mathcal{H}_i = \{\sigma(\mathbf{W} \cdot h) : h \in \mathcal{H}_{i-1}, \|\mathbf{W}\|_2 \leq B, \|\mathbf{W}^T\|_{2,1} \leq B_{2,1}\|\mathbf{W}\|_2)\}$. Let σ be L-Lipschitz. The Rademacher complexity bounds of Corollary 7 can be used to provide generalization bounds for neural networks. By Bartlett et al. (2017), the following upper bound holds for $\mathcal{H}_d$:

$$\widehat{\mathfrak{R}}_m^{\max}(\mathcal{H}) = \widetilde{O}\left(\frac{d^{3/2}BB_{2,1}}{\rho\sqrt{m}} \cdot (BL)^d\right).$$

Plugging in this upper bound in the bound of Corollary 7 leads to the following:

$$R(h) \leq \widehat{R}_S^\rho(h) + 2\sqrt{\widehat{R}_S^\rho(h)\,\beta_m} + \beta_m, \tag{4}$$

where $\beta_m = \widetilde{O}\left(\frac{d^3B^2B_{2,1}^2}{\rho^2 m} \cdot (BL)^{2d}\right)$. In comparison, the best existing neural network bounds by Bartlett et al. (2017, Theorem 1.1) is

$$R(h) \leq \widehat{R}_S^\rho(h) + c'\sqrt{\beta'_m}, \tag{5}$$

where c' is a universal constant and β'_m is the empirical Rademacher complexity. The margin bound (4) has the benefit of a more favorable dependency on the empirical margin loss than (5), which can be significant when that empirical term is small. On other hand, the empirical Rademacher complexity of (5) is more favorable than its counterpart in (4). A similar analysis can be used to derive relative margin bounds for ensembles of predictors or neural networks families (see Appendix F.2) as well as many other function classes.

6. Conclusion

Margin bounds are the most appropriate tools for the analysis of generalization in classification problems since they are more "optimistic" and typically not dimension-dependent. They have been used successfully to analyze the generalization properties of linear classifiers with Gaussian kernels, that of AdaBoost, and more recently that of neural networks. The finer margin guarantees we presented provide a more powerful tool for such analyses. Our relative margin bounds can further be used to derive guarantees for a variety of hypothesis sets and in a variety of applications. In particular, as illustrated in Appendix F.2, these bounds can help derive more favorable margin-based learning bounds for different families of neural networks, which has been the topic of several recent research publications. They may also serve as a useful tool in the analysis of scenarios such as active learning and the design of new algorithms.

References

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

M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207 – 217, 1993.

H. Bao, C. Scott, and M. Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *Conference on Learning Theory*, pages 408–451. PMLR, 2020.

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.

P. L. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In *Advances in Kernel Methods: Support Vector Learning*. MIT Press, 1998.

P. L. Bartlett, O. Bousquet, S. Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

P. L. Bartlett, D. J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Proceedings of NIPS*, pages 6240–6249, 2017.

C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3), 1995.

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

C. Cortes, M. Mohri, and U. Syed. Deep boosting. In *Proceedings of ICML*, pages 1179–1187, 2014.

C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, and S. Yang. Adanet: Adaptive structural learning of artificial neural networks. In *Proceedings of ICML*, pages 874–883, 2017.

C. Cortes, S. Greenberg, and M. Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *Ann. Math. Artif. Intell.*, 85(1):45–70, 2019.

R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2 – 142, 1984.

R. M. Dudley. Universal Donsker classes and metric entropy. *Annals of Probability*, 14(4):1306 – 1326, 1987.

D. J. Foster, S. Greenberg, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Hypothesis set stability and generalization. In *Proceedings of NeurIPS*, pages 6729–6739, 2019.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer System Sciences*, 55(1):119–139, 1997.

W. Gao and Z.-H. Zhou. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18, 2013.

S. Greenberg and M. Mohri. Tight lower bound on the probability of a binomial exceeding its expectation. *Statistics and Probability Letters*, 86:91–98, 2013.

A. Grønlund, L. Kamma, and K. G. Larsen. Near-tight margin-based generalization bounds for support vector machines. In *International Conference on Machine Learning*, pages 3779–3788. PMLR, 2020.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.

S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of NIPS*, pages 793–800, 2008.

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

V. Kuznetsov, M. Mohri, and U. Syed. Multi-class deep boosting. In *Proceedings of NIPS*, pages 2501–2509, 2014.

P. M. Long and H. Sedghi. Generalization bounds for deep convolutional neural networks. In *Proceedings of ICLR*, 2020.

D. McAllester. Simplified PAC-bayesian margin bounds. In *Learning theory and Kernel machines*, pages 203–215. Springer, 2003.

R. Meir and T. Zhang. Generalization Error Bounds for Bayesian Mixture Algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Proceedings of COLT*, pages 1376–1401, 2015.

D. Pollard. *Convergence of Stochastic Processess*. Springer, New York, 1984.

D. Pollard. Asymptotics via empirical processes. *Statistical Science*, 4(4):341 – 366, 1989.

R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of ICML*, pages 322–330, 1997.

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

N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Proceedings of NIPS*, pages 2199–2207, 2010.

B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Proceedings of NIPS*, 2003.

R. van Handel. *Probability in High Dimension, APC 550 Lecture Notes*. Princeton University, 2016.

V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data, second edition*. Springer, Berlin, 2006.

T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.