Rademacher Complexity Margin Bounds for Learning with a Large Number of Classes

Vitaly Kuznetsov VITALY@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, 251 Mercer street, New York, NY, 10012

Mehryar Mohri MOHRI@CS.NYU.EDU
Courant Institute and Google Research, 251 Mercer street, New York, NY, 10012

Umar Syed USYED@GOOGLE.COM
Google Research, 76 Ninth Avenue, New York, NY, 10011

Abstract

This paper presents improved Rademacher complexity margin bounds that scale linearly with the number of classes as opposed to the quadratic dependence of existing Rademacher complexity margin-based learning guarantees. We further use this result to prove a novel generalization bound for multi-class classifier ensembles that depends only on the Rademacher complexity of the hypothesis classes to which the classifiers in the ensemble belong.

1. Introduction

Multi-class classification is one of the central problems in machine learning. Given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ drawn i.i.d. from some unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{1, \ldots, c\}$, the objective of the learner consists of finding a hypothesis h that admits a small expected loss, h being selected out of some hypothesis class H. The expected loss is given by $\mathrm{E}_{(X,Y)\sim\mathcal{D}}[L(h(X), Y)]$, where L is a loss function, typically chosen to be the zero-one loss defined by $L(y', y) = 1_{y' \neq y}$.

A common approach to multi-class classification consists of learning a scoring function $f \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that assigns a score $f(x, y)$ to pair made of an input point $x \in \mathcal{X}$ and a candidate label y. The label predicted for x is the one with the highest score:

$$h(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} f(x, y).$$

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

The difference between the score of the correct label and that of the runner-up is the *margin* achieved for that example. The fraction of sample points with margin less than a specified constant ρ is the empirical margin loss of h. These quantities play a critical role in an algorithm-agnostic analysis of generalization in the multi-class setting based on data-dependent complexity measures such as Rademacher complexity. In particular, (Koltchinskii & Panchenko, 2002) showed that with high probability, uniformly over hypothesis set,

$$R(h) \leq \widehat{R}_{S,\rho}(h) + \frac{2c^2}{\rho}\mathfrak{R}_m(\Pi_1(G)) + O\left(\frac{1}{\sqrt{m}}\right),$$

where $R(h)$ is the generalization error of hypothesis h, $\widehat{R}_{S,\rho}(h)$ its empirical margin loss, and $\mathfrak{R}_m(\Pi_1(G))$ the Rademacher complexity of the family of loss functions $\Pi_1(G)$ associated to H, which is defined precisely below.

This bound is pessimistic and suggests that learning with an extremely large number of classes may not be possible. Indeed, it is well known that for certain classes of commonly used hypotheses, including linear and kernel-based ones, $\mathfrak{R}_m(\Pi_1(G)) \leq O(\frac{1}{\sqrt{m}})$. Therefore, for learning to occur we will need m to be on the order of at least $c^{4+\epsilon}/\rho$, for some $\epsilon > 0$. In some modern machine learning tasks such as speech recognition and image classification, c is often greater than 10^4. The bound above suggests that even for extremely favorable margin values of the order 10^3, a sample required for learning has to be in the order of at least 10^{13}. However, empirical results in speech recognition and image categorization suggest that it is possible to learn with much fewer samples. This result is also pessimistic in terms of computational complexity since storing and processing 10^{13} sample points may not be feasible. In this paper, we show that this bound can be improved to scale linearly with the number of classes.

We further consider convex ensembles of classification models. Ensemble methods are general techniques in machine learning for combining several multi-class classification hypothesis to further improve accuracy. Learning a linear combination of base classifiers, or a classifier ensemble, is one of the oldest and most powerful ideas in machine learning. Boosting (Freund & Schapire, 1997) — also known as forward stagewise additive modeling (Friedman et al., 1998) — is a widely used meta-algorithm for ensemble learning. In the boosting approach, the ensemble's misclassification error is replaced by a convex upper bound, called the surrogate loss. The algorithm greedily minimizes the surrogate loss by augmenting the ensemble with a classifier (or adjusting the weight of a classifier already in the ensemble) at each of iteration.

One of the main advantages of boosting is that, because it is a stagewise procedure, one can efficiently learn a classifier ensemble in which each classifier belongs to a large (and potentially infinite) base hypothesis class, provided that one has an efficient algorithm for learning good base classifiers. For example, decision trees are commonly used as the base hypothesis class. In contrast, generalization bounds for classifier ensembles tend to increase with the complexity of the base hypothesis class (Schapire et al., 1997), and indeed boosting has been observed to overfit in practice (Grove & Schuurmans, 1998; Schapire, 1999; Dietterich, 2000; Rätsch et al., 2001b).

One way to address overfitting in a boosted ensemble is to regularize the weights of the classifiers. Standard regularization penalizes all the weights in the ensemble equally (Rätsch et al., 2001a; Duchi & Singer, 2009), but in some cases it seems they should be penalized unequally. For example, in an ensemble of decision trees, deeper decision trees should have a larger regularization penalty than shallower ones. Based on this idea, we present a novel generalization guarantee for multi-class classifier ensembles that depends only on the Rademacher complexity of the hypothesis classes to which the classifiers in the ensemble belong.

(Cortes et al., 2014) developed this idea in an algorithm called *DeepBoost*, a boosting algorithm where the decision in each iteration of which classifier to add to the ensemble, and the weight assigned to that classifier, depends in part on the complexity of the hypothesis class to which it belongs. One interpretation of DeepBoost is that it applies the principle of structural risk minimization to each iteration of boosting. (Kuznetsov et al., 2014) extended these ideas to the multi-class setting.

The rest of this paper is organized as follows. In Section 2 we present and prove our improved Rademacher complexity margin bounds that scale linearly with the number of classes. In Section 3 we use this result to prove a novel generalization bound for multi-class classifier ensembles that depends only on the Rademacher complexity of the hypothesis classes to which the classifiers in the ensemble belong. We conclude with some final remarks in Section 4.

2. Multi-class margin bounds

In this section, we present our improved data-dependent learning bound in the multi-class setting. Let $\mathcal{X}$ denote the input space. We denote by $\mathcal{Y} = \{1, \ldots, c\}$ a set of c classes, which, for convenience, we index by an integer in $[1, c]$. The label associated by a hypothesis $f \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to $x \in \mathcal{X}$ is given by $\operatorname{argmax}_{y \in \mathcal{Y}} f(x, y)$. The margin $\rho_f(x, y)$ of the function f for a labeled example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is defined by

$$\rho_f(x, y) = f(x, y) - \max_{y' \neq y} f(x, y'). \tag{1}$$

Thus, f misclassifies (x, y) iff $\rho_f(x, y) \leq 0$. We assume that training and test points are drawn i.i.d. according to some distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and denote by $S = ((x_1, y_1), \ldots, (x_m, y_m))$ a training sample of size m drawn according to $\mathcal{D}^m$. For any $\rho > 0$, the generalization error $R(f)$, its ρ-margin error $R_\rho(f)$ and its empirical margin error are defined as follows:

$$R(f) = \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}[1_{\rho_f(x,y)\leq 0}],$$
$$R_\rho(f) = \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}[1_{\rho_f(x,y)\leq \rho}],$$
$$\widehat{R}_{S,\rho}(f) = \mathop{\mathrm{E}}_{(x,y)\sim S}[1_{\rho_f(x,y)\leq \rho}],$$

where the notation $(x, y) \sim S$ indicates that (x, y) is drawn according to the empirical distribution defined by S. For any family of hypotheses G mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, we define $\Pi_1(G)$ by

$$\Pi_1(G) = \{x \mapsto h(x, y) \colon y \in \mathcal{Y}, h \in G\}. \tag{2}$$

The following result due to (Koltchinskii & Panchenko, 2002) is a well known margin bound for the multi-class setting.

Theorem 1. *Let G be a family of hypotheses mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, with $\mathcal{Y} = \{1, \ldots, c\}$. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta > 0$, the following bound holds for all $g \in G$:*

$$R(g) \leq \widehat{R}_{S,\rho}(g) + \frac{2c^2}{\rho}\mathfrak{R}_m(\Pi_1(G)) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

where $\Pi_1(G) = \{(x, y) \mapsto g(x, y) : y \in \mathcal{Y}, g \in G\}$.

As discussed in the introduction, the bound of Theorem 1 is pessimistic and suggests that learning with an extremely large number of classes may not be possible. The following theorem presents our margin learning guarantees for

multi-class classification with a large number of classes that scales linearly with the number of classes, as opposed to the quadratic dependency of Theorem 1.

Theorem 2. *Let G be a family of hypotheses mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, with $\mathcal{Y} = \{1, \ldots, c\}$. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta > 0$, the following bound holds for all $g \in G$:*

$$R(g) \leq \widehat{R}_{S,\rho}(g) + \frac{4c}{\rho}\mathfrak{R}_m(\Pi_1(G)) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

where $\Pi_1(G) = \{(x, y) \mapsto g(x, y) : y \in \mathcal{Y}, g \in G\}$.

Note that the bound of Theorem 2 is strictly better than that of Theorem 1 for all $c > 2$. The bound of Theorem 2 is more optimistic both in terms of computation resources and statistical hardness of the problem. To the best of our knowledge, it is an open problem if the dependence on the number of classes can be further improved in general, that is for arbitrary hypothesis sets.

Proof. We will need the following definition for this proof:

$$\rho_g(x, y) = \min_{y' \neq y}(g(x, y) - g(x, y'))$$
$$\rho_{\theta,g}(x, y) = \min_{y'}(g(x, y) - g(x, y') + \theta 1_{y'=y}),$$

where $\theta > 0$ is an arbitrary constant. Observe that $\mathrm{E}\, 1_{\rho_g(x,y)\leq 0} \leq \mathrm{E}\, 1_{\rho_{\theta,g}(x,y)\leq 0}$. To verify this claim it suffices to check that $1_{\rho_g(x,y)\leq 0} \leq 1_{\rho_{\theta,g}(x,y)\leq 0}$, which is equivalent to the following statement: if $\rho_g(x, y) \leq 0$ then $\rho_{\theta,g}(x, y) \leq 0$. Indeed, this follows from the following bound:

$$\begin{aligned}\rho_{\theta,g}(x, y) &= \min_{y'} \big(g(x, y) - g(x, y') + \theta 1_{y'=y}\big) \\ &\leq \min_{y' \neq y} \big(g(x, y) - g(x, y') + \theta 1_{y'=y}\big) \\ &= \min_{y' \neq y} \big(g(x, y) - g(x, y')\big) = \rho_g(x, y),\end{aligned}$$

where the inequality follows from taking the minimum over a smaller set.

Let Φ_ρ be the margin loss function defined for all $u \in \mathbb{R}$ by $\Phi_\rho(u) = 1_{u\leq 0} + (1 - \frac{u}{\rho})1_{0<u\leq\rho}$. We also let $\widetilde{G} = \{(x, y) \mapsto \rho_{\theta,g}(x, y) \colon g \in G\}$ and $\widetilde{\mathcal{G}} = \{\Phi_\rho \circ \widetilde{g} \colon \widetilde{g} \in \widetilde{G}\}$. By the standard Rademacher complexity bound (Koltchinskii & Panchenko, 2002; Mohri et al., 2012), for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $g \in G$:

$$R(g) \leq \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(\rho_{\theta,g}(x_i, y_i)) + 2\mathfrak{R}_m(\widetilde{\mathcal{G}}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Fixing $\theta = 2\rho$, we observe that $\Phi_\rho(\rho_{\theta,g}(x_i, y_i)) = \Phi_\rho(\rho_g(x_i, y_i)) \leq 1_{\rho_g(x_i,y_i)\leq\rho}$. Indeed, either $\rho_{\theta,g}(x_i, y_i) = \rho_g(x_i, y_i)$ or $\rho_{\theta,g}(x_i, y_i) = 2\rho \leq \rho_g(x_i, y_i)$, which implies the desired result. Talagrand's lemma (Ledoux & Talagrand, 1991; Mohri et al., 2012) yields $\mathfrak{R}_m(\widetilde{\mathcal{G}}) \leq \frac{1}{\rho}\mathfrak{R}_m(\widetilde{G})$ since Φ_ρ is a $\frac{1}{\rho}$-Lipschitz function. Therefore, for any $\delta > 0$, with probability at least $1 - \delta$, for all $g \in G$:

$$R(g) \leq R_{S,\rho}(g) + \frac{2}{\rho}\mathfrak{R}_m(\widetilde{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

and to complete the proof it suffices to show that $\mathfrak{R}_m(\widetilde{G}) \leq 2c\mathfrak{R}_m(\Pi_1(G))$.

Here, $\mathfrak{R}_m(\widetilde{G})$ can be upper-bounded as follows:

$$\begin{aligned}\mathfrak{R}_m(\widetilde{G}) &= \frac{1}{m}\mathop{\mathrm{E}}_{S,\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sigma_i(g(x_i, y_i) \\ &\qquad - \max_y(g(x_i, y) - 2\rho 1_{y=y_i}))\Big] \\ &\leq \frac{1}{m}\mathop{\mathrm{E}}_{S,\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sigma_i g(x_i, y_i)\Big] \\ &\qquad + \frac{1}{m}\mathop{\mathrm{E}}_{S,\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sigma_i \max_y(g(x_i, y) - 2\rho 1_{y=y_i})\Big].\end{aligned}$$

Now we bound the second term above. Observe that

$$\begin{aligned}&\frac{1}{m}\mathop{\mathrm{E}}_{\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sigma_i g(x_i, y_i)\Big] \\ &= \frac{1}{m}\mathop{\mathrm{E}}_{\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sum_{y\in\mathcal{Y}}\sigma_i g(x_i, y)1_{y_i=y}\Big] \\ &\leq \frac{1}{m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sigma_i g(x_i, y)1_{y_i=y}\Big] \\ &= \sum_{y\in\mathcal{Y}}\frac{1}{m}\mathop{\mathrm{E}}_{\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sigma_i g(x_i, y)\Big(\frac{\epsilon_i}{2} + \frac{1}{2}\Big)\Big],\end{aligned}$$

where $\epsilon_i = 2 \cdot 1_{y_i=y} - 1$. Since $\epsilon_i \in \{-1, +1\}$, σ_i and $\sigma_i\epsilon_i$ admit the same distribution and, for any $y \in \mathcal{Y}$, each of the terms of the right-hand side can be bounded as follows:

$$\begin{aligned}&\frac{1}{m}\mathop{\mathrm{E}}_{\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sigma_i g(x_i, y)\Big(\frac{\epsilon_i}{2} + \frac{1}{2}\Big)\Big] \\ &\qquad \leq \frac{1}{2m}\mathop{\mathrm{E}}_{\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sigma_i\epsilon_i g(x_i, y)\Big] \\ &\qquad\quad + \frac{1}{2m}\mathop{\mathrm{E}}_{\sigma}\Big[\sup_{g\in G}\sum_{i=1}^{m}\sigma_i g(x_i, y)\Big] \\ &\qquad \leq \widehat{\mathfrak{R}}_m(\Pi_1(G)).\end{aligned}$$

Thus, we can write $\frac{1}{m}\operatorname{E}_{S,\sigma}\left[\sup_{g\in G}\sum_{i=1}^m \sigma_i g(x_i,y_i)\right] \leq c\,\mathfrak{R}_m(\Pi_1(G))$. To bound the second term, we first apply Lemma 8.1 of (Mohri et al., 2012) that immediately yields that

$$\frac{1}{m}\operatorname*{E}_{S,\sigma}\left[\sup_{g\in G}\sum_{i=1}^m \sigma_i \max_y (g(x_i,y)-2\rho 1_{y=y_i})\right] \leq \sum_{y\in\mathcal{Y}}\frac{1}{m}\operatorname*{E}_{S,\sigma}\left[\sup_{g\in G}\sum_{i=1}^m \sigma_i (g(x_i,y)-2\rho 1_{y=y_i})\right]$$

and since Rademacher variables are mean zero, we observe that

$$\begin{aligned}&\operatorname*{E}_{S,\sigma}\left[\sup_{g\in G}\sum_{i=1}^m \sigma_i (g(x_i,y)-2\rho 1_{y=y_i})\right]\\ &= \operatorname*{E}_{S,\sigma}\left[\sup_{g\in G}\left(\sum_{i=1}^m \sigma_i g(x_i,y)\right)-2\rho\sum_{i=1}^m \sigma_i 1_{y=y_i}\right]\\ &= \operatorname*{E}_{S,\sigma}\left[\sup_{g\in G}\sum_{i=1}^m \sigma_i g(x_i,y)\right] \leq \mathfrak{R}_m(\Pi_1(G))\end{aligned}$$

which completes the proof. □

3. Multi-class data-dependent learning guarantee for convex ensembles

We consider p families $H_1,\ldots,H_p$ of functions mapping from $\mathcal{X}\times\mathcal{Y}$ to $[0,1]$ and the ensemble family $\mathcal{F} = \operatorname{conv}(\bigcup_{k=1}^p H_k)$, that is the family of functions f of the form $f=\sum_{t=1}^T \alpha_t h_t$, where $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_T)$ is in the simplex Δ and where, for each $t\in[1,T]$, h_t is in H_{k_t} for some $k_t\in[1,p]$.

The following theorem gives a margin-based Rademacher complexity bound for learning with ensembles of base classifiers with multiple hypothesis sets. As with other Rademacher complexity learning guarantees, our bound is data-dependent, which is an important and favorable characteristic of our results.

Theorem 3. *Assume $p>1$ and let $H_1,\ldots,H_p$ be p families of functions mapping from $\mathcal{X}\times\mathcal{Y}$ to $[0,1]$. Fix $\rho>0$. Then, for any $\delta>0$, with probability at least $1-\delta$ over the choice of a sample S of size m drawn i.i.d. according to $\mathcal{D}$, the following inequality holds for all $f=\sum_{t=1}^T \alpha_t h_t\in\mathcal{F}$:*

$$R(f) \leq \widehat{R}_{S,\rho}(f) + \frac{8c}{\rho}\sum_{t=1}^T \alpha_t \mathfrak{R}_m(\Pi_1(H_{k_t})) + \frac{2}{c\rho}\sqrt{\frac{\log p}{m}} + \sqrt{\left\lceil \frac{4}{\rho^2}\log\left(\frac{c^2\rho^2 m}{4\log p}\right)\right\rceil \frac{\log p}{m} + \frac{\log\frac{2}{\delta}}{2m}},$$

Thus, $R(f) \leq \widehat{R}_{S,\rho}(f) + \frac{8c}{\rho}\sum_{t=1}^T \alpha_t \mathfrak{R}_m(H_{k_t}) + O\left(\sqrt{\frac{\log p}{\rho^2 m}\log\left[\frac{\rho^2 c^2 m}{4\log p}\right]}\right)$.

Before we present the proof of this result we discuss some of its consequences. For $p=1$, that is for the special case of a single hypothesis set, this bound to the bound of Theorem 2. However, the main remarkable benefit of this learning bound is that its complexity term admits an explicit dependency on the mixture coefficients α_t. It is a weighted average of Rademacher complexities with mixture weights α_t, $t\in[1,T]$. Thus, the second term of the bound suggests that, while some hypothesis sets H_k used for learning could have a large Rademacher complexity, this may not negatively affect generalization if the corresponding total mixture weight (sum of α_ts corresponding to that hypothesis set) is relatively small. Using such potentially complex families could help achieve a better margin on the training sample.

The theorem cannot be proven via the standard Rademacher complexity analysis of Koltchinskii & Panchenko (2002) since the complexity term of the bound would then be $\mathfrak{R}_m(\operatorname{conv}(\bigcup_{k=1}^p H_k)) = \mathfrak{R}_m(\bigcup_{k=1}^p H_k)$ which does not admit an explicit dependency on the mixture weights and is lower bounded by $\sum_{t=1}^T \alpha_t \mathfrak{R}_m(H_{k_t})$. Thus, the theorem provides a finer learning bound than the one obtained via a standard Rademacher complexity analysis.

Our proof makes use of Theorem 2 and a proof technique used in (Schapire et al., 1997).

Proof. For a fixed $\mathbf{h}=(h_1,\ldots,h_T)$, any $\boldsymbol{\alpha}$ in the probability simplex Δ defines a distribution over $\{h_1,\ldots,h_T\}$. Sampling from $\{h_1,\ldots,h_T\}$ according to $\boldsymbol{\alpha}$ and averaging leads to functions g of the form $g=\frac{1}{n}\sum_{i=1}^T n_t h_t$ for some $\mathbf{n}=(n_1,\ldots,n_T)$, with $\sum_{t=1}^T n_t = n$, and $h_t\in H_{k_t}$.

For any $\mathbf{N}=(N_1,\ldots,N_p)$ with $|\mathbf{N}|=n$, we consider the family of functions

$$G_{\mathcal{F},\mathbf{N}} = \left\{\frac{1}{n}\sum_{k=1}^p\sum_{j=1}^{N_k} h_{k,j} \mid \forall (k,j)\in[p]\times[N_k], h_{k,j}\in H_k\right\},$$

and the union of all such families $G_{\mathcal{F},n} = \bigcup_{|\mathbf{N}|=n} G_{\mathcal{F},\mathbf{N}}$. Fix $\rho>0$. For a fixed $\mathbf{N}$, the Rademacher complexity of $\Pi_1(G_{\mathcal{F},\mathbf{N}})$ can be bounded as follows for any $m\geq 1$: $\mathfrak{R}_m(\Pi_1(G_{\mathcal{F},\mathbf{N}})) \leq \frac{1}{n}\sum_{k=1}^p N_k\,\mathfrak{R}_m(\Pi_1(H_k))$. Thus, by Theorem 2, the following multi-class margin-based Rademacher complexity bound holds. For any $\delta>0$, with probability at least $1-\delta$, for all $g\in G_{\mathcal{F},\mathbf{N}}$,

$$R_\rho(g) - \widehat{R}_{S,\rho}(g) \leq \frac{1}{n}\frac{4c}{\rho}\sum_{k=1}^p N_k\,\mathfrak{R}_m(\Pi_1(H_k)) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Since there are at most p^n possible p-tuples $\mathbf{N}$ with $|\mathbf{N}| =$

n,[1] by the union bound, for any $\delta > 0$, with probability at least $1-\delta$, for all $g \in G_{\mathcal{F},n}$, we can write

$$R_\rho(g) - \widehat{R}_{S,\rho}(g) \leq \frac{1}{n}\frac{4c}{\rho}\sum_{k=1}^{p} N_k \mathfrak{R}_m(\Pi_1(H_k)) + \sqrt{\frac{\log \frac{p^n}{\delta}}{2m}}.$$

Thus, with probability at least $1-\delta$, for all functions $g = \frac{1}{n}\sum_{i=1}^{T} n_t h_t$ with $h_t \in H_{k_t}$, the following inequality holds

$$R_\rho(g) - \widehat{R}_{S,\rho}(g) \leq \frac{1}{n}\frac{4c}{\rho}\sum_{k=1}^{p}\sum_{t:k_t=k} n_t \mathfrak{R}_m(\Pi_1(H_{k_t})) + \sqrt{\frac{\log \frac{p^n}{\delta}}{2m}}.$$

Taking the expectation with respect to $\boldsymbol{\alpha}$ and using $\mathrm{E}_{\boldsymbol{\alpha}}[n_t/n] = \alpha_t$, we obtain that for any $\delta > 0$, with probability at least $1-\delta$, for all g, we can write

$$\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_\rho(g) - \widehat{R}_{S,\rho}(g)] \leq \frac{4c}{\rho}\sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(\Pi_1(H_{k_t})) + \sqrt{\frac{\log \frac{p^n}{\delta}}{2m}}.$$

Fix $n \geq 1$. Then, for any $\delta_n > 0$, with probability at least $1-\delta_n$,

$$\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)] \leq \frac{8c}{\rho}\sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(\Pi_1(H_{k_t})) + \sqrt{\frac{\log \frac{p^n}{\delta_n}}{2m}}.$$

Choose $\delta_n = \frac{\delta}{2p^{n-1}}$ for some $\delta > 0$, then for $p \geq 2$, $\sum_{n\geq 1}\delta_n = \frac{\delta}{2(1-1/p)} \leq \delta$. Thus, for any $\delta > 0$ and any $n \geq 1$, with probability at least $1-\delta$, the following holds for all g:

$$\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)] \leq \frac{8c}{\rho}\sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(\Pi_1(H_{k_t})) + \sqrt{\frac{\log \frac{2p^{2n-1}}{\delta}}{2m}}. \tag{3}$$

Now, for any $f = \sum_{t=1}^{T}\alpha_t h_t \in \mathcal{F}$ and any $g = \frac{1}{n}\sum_{i=1}^{T} n_t h_t$, we can upper-bound $R(f) = \Pr_{(x,y)\sim\mathcal{D}}[\rho_f(x,y) \leq 0]$, the generalization error of f, as follows:

$$\begin{aligned}
R(f) &= \Pr_{(x,y)\sim\mathcal{D}}[\rho_f(x,y) - \rho_g(x,y) + \rho_g(x,y) \leq 0] \\
&\leq \Pr[\rho_f(x,y) - \rho_g(x,y) < -\rho/2] + \Pr[\rho_g(x,y) \leq \rho/2] \\
&= \Pr[\rho_f(x,y) - \rho_g(x,y) < -\rho/2] + R_{\rho/2}(g).
\end{aligned}$$

We can also write

$$\widehat{R}_{\rho/2}(g) = \widehat{R}_{S,\rho/2}(g - f + f) \leq \Pr_{(x,y)\sim S}[\rho_g(x,y) - \rho_f(x,y) < -\rho/2] + \widehat{R}_{S,\rho}(f).$$

Combining these inequalities yields

$$\begin{aligned}
&\Pr_{(x,y)\sim\mathcal{D}}[\rho_f(x,y) \leq 0] - \widehat{R}_{S,\rho}(f) \\
&\leq \Pr_{(x,y)\sim\mathcal{D}}[\rho_f(x,y) - \rho_g(x,y) < -\rho/2] + \Pr_{(x,y)\sim S}[\rho_g(x,y) - \rho_f(x,y) < -\rho/2] \\
&\quad + R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g).
\end{aligned}$$

Taking the expectation with respect to $\boldsymbol{\alpha}$ yields

$$\begin{aligned}
R(f) - \widehat{R}_{S,\rho}(f) \leq{}& \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D},\boldsymbol{\alpha}}[1_{\rho_f(x,y)-\rho_g(x,y)<-\rho/2}] \\
&+ \mathop{\mathrm{E}}_{(x,y)\sim S,\boldsymbol{\alpha}}[1_{\rho_g(x,y)-\rho_f(x,y)<-\rho/2}] \\
&+ \mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)].
\end{aligned} \tag{4}$$

Fix (x,y) and for any function $\varphi\colon \mathcal{X}\times\mathcal{Y} \to [0,1]$ define y'_φ as follows: $y'_\varphi = \operatorname{argmax}_{y'\neq y}\varphi(x,y)$. For any g, by definition of ρ_g, we can write $\rho_g(x,y) \leq g(x,y) - g(x,y'_f)$. In light of this inequality and Hoeffding's bound, the following holds:

$$\begin{aligned}
&\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[1_{\rho_f(x,y)-\rho_g(x,y)<-\rho/2}] \\
&= \Pr_{\boldsymbol{\alpha}}[\rho_f(x,y) - \rho_g(x,y) < -\rho/2] \\
&\leq \Pr_{\boldsymbol{\alpha}}\Big[\big(f(x,y) - f(x,y'_f)\big) - \big(g(x,y) - g(x,y'_f)\big) < -\rho/2\Big] \\
&\leq e^{-n\rho^2/8}.
\end{aligned}$$

Similarly, for any g, we can write $\rho_f(x,y) \leq f(x,y) - f(x,y'_g)$. Using this inequality, the union bound and Hoeffding's bound, the other expectation term appearing on the right-hand side of (4) can be bounded as follows:

$$\begin{aligned}
&\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[1_{\rho_g(x,y)-\rho_f(x,y)<-\rho/2}] \\
&= \Pr_{\boldsymbol{\alpha}}[\rho_g(x,y) - \rho_f(x,y) < -\rho/2] \\
&\leq \Pr_{\boldsymbol{\alpha}}\Big[\big(g(x,y) - g(x,y'_g)\big) - \big(f(x,y) - f(x,y'_g)\big) < -\rho/2\Big] \\
&\leq \sum_{y'\neq y}\Pr_{\boldsymbol{\alpha}}\Big[\big(g(x,y) - g(x,y')\big) - \big(f(x,y) - f(x,y')\big) < -\rho/2\Big] \\
&\leq (c-1)e^{-n\rho^2/8}.
\end{aligned}$$

[1] The number $S(p,n)$ of p-tuples $\mathbf{N}$ with $|\mathbf{N}| = n$ is known to be precisely $\binom{p+n-1}{p-1}$.

Thus, for any fixed $f \in \mathcal{F}$, we can write

$$R(f) - \widehat{R}_{S,\rho}(f) \leq ce^{-n\rho^2/8} + \mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)].$$

Therefore, the following inequality holds:

$$\sup_{f\in\mathcal{F}} R(f) - \widehat{R}_{S,\rho}(f) \leq ce^{-n\rho^2/8} + \sup_g \mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)],$$

and, in view of (3), for any $\delta > 0$ and any $n \geq 1$, with probability at least $1-\delta$, the following holds for all $f \in \mathcal{F}$:

$$R(f) - \widehat{R}_{S,\rho}(f) \leq \frac{8c}{\rho} \sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(\Pi_1(H_{k_t})) + ce^{-\frac{n\rho^2}{8}} + \sqrt{\frac{(2n-1)\log p + \log\frac{2}{\delta}}{2m}}.$$

Choosing $n = \left\lceil \frac{4}{\rho^2} \log\left(\frac{c^2\rho^2 m}{4\log p}\right)\right\rceil$ yields the following inequality:[2]

$$R(f) - \widehat{R}_{S,\rho}(f) \leq \frac{8c}{\rho} \sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(\Pi_1(H_{k_t})) + \frac{2}{c\rho}\sqrt{\frac{\log p}{m}} + \sqrt{\left\lceil \frac{4}{\rho^2} \log\left(\frac{c^2\rho^2 m}{4\log p}\right)\right\rceil \frac{\log p}{m} + \frac{\log\frac{2}{\delta}}{2m}},$$

and concludes the proof. □

4. Conclusion

We presented improved Rademacher complexity margin bounds that scale linearly with the number of classes, as opposed to the quadratic dependency of the existing Rademacher complexity margin-based learning guarantees. Furthermore, we used this result to prove a novel generalization bound for multi-class classifier ensembles that depends only on the Rademacher complexity of the hypothesis classes to which the classifiers in the ensemble belong.

(Cortes et al., 2014) developed this idea in an algorithm called DeepBoost, a boosting algorithm where the decision at each iteration of which classifier to add to the ensemble, and which weight to assign to that classifier, depends on the complexity of the hypothesis class to which it belongs. One interpretation of DeepBoost is that it applies the principle of structural risk minimization to each iteration of boosting. (Kuznetsov et al., 2014) extended these ideas to the multi-class setting.

[2] To select n we consider $f(n) = ce^{-nu} + \sqrt{nv}$, where $u = \rho^2/8$ and $v = \log p/m$. Taking the derivative of f, setting it to zero and solving for n, we obtain $n = -\frac{1}{2u} W_{-1}(-\frac{v}{2c^2u})$ where W_{-1} is the second branch of the Lambert function (inverse of $x \mapsto xe^x$). Using the bound $-\log x \leq -W_{-1}(-x) \leq 2\log x$ leads to the following choice of n: $n = \left\lceil -\frac{1}{2u}\log(\frac{v}{2c^2u})\right\rceil$.

References

Cortes, Corinna, Mohri, Mehryar, and Syed, Umar. Deep boosting. In *ICML*, pp. 1179 – 1187, 2014.

Dietterich, Thomas G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.

Duchi, John C. and Singer, Yoram. Boosting with structural sparsity. In *ICML*, pp. 38, 2009.

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

Friedman, Jerome H., Hastie, Trevor, and Tibshirani, Robert. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000, 1998.

Grove, Adam J and Schuurmans, Dale. Boosting in the limit: Maximizing the margin of learned ensembles. In *AAAI/IAAI*, pp. 692–699, 1998.

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

Kuznetsov, Vitaly, Mohri, Mehryar, and Syed, Umar. Multi-class deep boosting. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2501–2509. Curran Associates, Inc., 2014.

Ledoux, Michel and Talagrand, Michel. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet. *Foundations of Machine Learning*. The MIT Press, 2012.

Rätsch, Gunnar, Mika, Sebastian, and Warmuth, Manfred K. On the convergence of leveraging. In *NIPS*, pp. 487–494, 2001a.

Rätsch, Gunnar, Onoda, Takashi, and Müller, Klaus-Robert. Soft margins for AdaBoost. *Machine Learning*, 42(3):287–320, 2001b.

Schapire, Robert E. Theoretical views of boosting and applications. In *Proceedings of ALT 1999*, volume 1720 of *Lecture Notes in Computer Science*, pp. 13–25. Springer, 1999.

Schapire, Robert E., Freund, Yoav, Bartlett, Peter, and Lee, Wee Sun. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pp. 322–330, 1997.