Smoothness, Low-Noise and Fast Rates

Nathan Srebro nati@ttic.edu
Karthik Sridharan karthik@ttic.edu
Toyota Technological Institute at Chicago

Ambuj Tewari ambuj@cs.utexas.edu
Computer Science Dept., University of Texas at Austin

Abstract

We establish an excess risk bound of $\tilde{O}\left(H\mathcal{R}_n^2 + \sqrt{HL^*}\mathcal{R}_n\right)$ for ERM with an H-smooth loss function and a hypothesis class with Rademacher complexity $\mathcal{R}_n$, where L^* is the best risk achievable by the hypothesis class. For typical hypothesis classes where $\mathcal{R}_n = \sqrt{R/n}$, this translates to a learning rate of $\tilde{O}\left(RH/n\right)$ in the separable ($L^* = 0$) case and $\tilde{O}\left(RH/n + \sqrt{L^* RH/n}\right)$ more generally. We also provide similar guarantees for online and stochastic convex optimization of a smooth non-negative objective.

1 Introduction

Consider empirical risk minimization for a hypothesis class $\mathcal{H} = \{h : \mathcal{X} \to \mathbb{R}\}$ w.r.t. some non-negative loss function $\phi(t, y)$. That is, we would like to learn a predictor h with small risk $L\left(h\right) = \mathbb{E}\left[\phi(h(X), Y)\right]$ by minimizing the empirical risk $\hat{L}(h) = \frac{1}{n}\sum_{i=1}^{n}\phi(h(x_i), y_i)$ of an i.i.d. sample $(x_1, y_1), \ldots, (x_n, y_n)$.

Statistical guarantees on the excess risk are well understood for *parametric* (i.e. finite dimensional) hypothesis classes. More formally, these are hypothesis classes with finite VC-subgraph dimension [23] (aka pseudo-dimension). For such classes learning guarantees can be obtained for any bounded loss function (i.e. s.t. $|\phi| \le b < \infty$) and the relevant measure of complexity is the VC-subgraph dimension.

Alternatively, even for some non-parametric hypothesis classes (i.e. those with infinite VC-subgraph dimension), e.g. the class of low-norm linear predictors $\mathcal{H}_B = \{h_w : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle | \|\mathbf{w}\| \le B\}$, guarantees can be obtained in terms of *scale-sensitive* measures of complexity such as fat-shattering dimensions [1], covering numbers [23] or Rademacher complexity [2]. The classical statistical learning theory approach for obtaining learning guarantees for such scale-sensitive classes is to rely on the Lipschitz constant D of $\phi(t, y)$ w.r.t. t (i.e. bound on its derivative w.r.t. t). The excess risk can then be bounded as (in expectation over the sample):

$$L\left(\hat{h}\right) \le L^* + 2D\mathcal{R}_n(\mathcal{H}) = L^* + 2\sqrt{D^2\frac{R}{n}} \tag{1}$$

where $\hat{h} = \arg\min \hat{L}(h)$ is the empirical risk minimizer (ERM), $L^* = \inf_h L\left(h\right)$ is the approximation error, and $\mathcal{R}_n(\mathcal{H})$ is the Rademacher complexity, which typically scales as $\mathcal{R}_n(\mathcal{H}) = \sqrt{R/n}$. E.g. for ℓ_2-bounded linear predictors, $R = B^2 \sup \|X\|_2^2$.

In this paper we address two deficiencies of the guarantee (1). First, the bound applies only to loss functions with bounded derivative, like the hinge loss and logistic loss popular for classification, or the absolute-value (ℓ_1) loss for regression. It is not directly applicable to the squared loss $\phi(t, y) = \frac{1}{2}(t - y)^2$, for which the second derivative is bounded, but not the first. We could try to simply bound the derivative of the squared loss in terms of a bound on the magnitude of $h(x)$, but e.g. for norm-bounded linear predictors $\mathcal{H}_B$ this results in a very disappointing excess risk bound of the form $O(\sqrt{B^4(\max\|X\|)^4/n})$. One aim of this paper is to provide clean bounds on the excess risk for smooth loss functions such as the squared loss with a bounded second, rather then first, derivative.

The second deficiency of (1) is the dependence on $1/\sqrt{n}$. The dependence on $1/\sqrt{n}$ might be unavoidable in general. But at least for finite dimensional (parametric) classes, we know it can be improved to a $1/n$ rate when the distribution is separable, i.e. when there exists $h \in \mathcal{H}$ with $L(h) = 0$ and so $L^* = 0$. In particular, if $\mathcal{H}$ is a class of bounded functions with VC-subgraph-dimension d (e.g. d-dimensional linear predictors), then in expectation over sample [22]:

$$L\left(\hat{h}\right) \leq L^* + O\left(\frac{dD \log n}{n} + \sqrt{\frac{dDL^* \log n}{n}}\right) \tag{2}$$

The $\sqrt{1/n}$ term disappears in the separable case, and we get a graceful degradation between the $\sqrt{1/n}$ rate to the $1/n$ rate for separable case. Could we get a $1/n$ separable rate, and such a graceful degradation, in non-parametric case?

As we will show, the two deficiencies are actually related. For non-parametric classes, and non-smooth Lipschitz loss, such as the hinge-loss, the excess risk might scale as $\sqrt{1/n}$ and not $1/n$, *even in the separable case*. However, for H-smooth non-negative loss functions, where the second derivative of $\phi(t, y)$ w.r.t. t is bounded by H, a $1/n$ separable rate *is* possible. In Section 2 we obtain the following bound on the excess risk (up to logarithmic factors):

$$L\left(\hat{h}\right) \leq L^* + \tilde{O}\left(H\mathcal{R}_n^2(\mathcal{H}) + \sqrt{HL^*}\mathcal{R}_n(\mathcal{H})\right) = L^* + \tilde{O}\left(\frac{HR}{n} + \sqrt{\frac{HRL^*}{n}}\right) \;\leq\; 2L^* + \tilde{O}\left(\frac{HR}{n}\right). \tag{3}$$

In particular, for ℓ_2-norm-bounded linear predictors $\mathcal{H}_B$ with $\sup \|X\|_2^2 \leq 1$, the excess risk is bounded by $\tilde{O}(HB^2/n + \sqrt{HB^2L^*/n})$. Another interesting distinction between parametric and non-parametric classes, is that even for the squared-loss, the bound (3) is tight and the non-separable rate of $1/\sqrt{n}$ is unavoidable. This is in contrast to the parametric (fine dimensional) case, where a rate of $1/n$ is always possible for the squared loss, regardless of the approximation error L^* [16]. The differences between parametric and scale-sensitive classes, and between non-smooth, smooth and strongly convex loss functions are discussed in Section 4 and summarized in Table 1.

The guarantees discussed thus far are general learning guarantees for the stochastic setting that rely only on the Rademacher complexity of the hypothesis class, and are phrased in terms of minimizing some scalar loss function. In Section 3 we consider also the online setting, in addition to the stochastic setting, and present similar guarantees for online and stochastic convex optimization [32, 24]. The guarantees of Section 3 match equation (3) for the special case of a convex loss function and norm-bounded linear predictors, but Section 3 capture a more general setting of optimizing an arbitrary non-negative convex objective, which we require to be smooth (there is no separate discussion of a "predictor" and a scalar loss function in Section 3). Results in Section 3 are expressed in terms of properties of the norm, rather then a measure of concentration like the Radamacher complexity as in (3) and Section 2. However, the online and stochastic convex optimization setting of Section 3 is also more restrictive, as we require the objective be convex (in Section 2 we make no assumption about the convexity of hypothesis class $\mathcal{H}$ nor the loss function ϕ).

Specifically, for a non-negative H-smooth *convex* objective, over a domain bounded by B, we prove that the average online regret (and excess risk of stochastic optimization) is bounded by $O(HB^2/n + \sqrt{HB^2L^*/n})$. Comparing with the bound of $O(\sqrt{D^2B^2/n})$ when the loss is D-Lipschitz rather then H-smooth [32, 21], we see the same relationship discussed above for ERM. Unlike the bound (3) for the ERM, the convex optimization bound avoids polylogarithmic factors. The results in Section 3 also generalize to smoothness and boundedness with respect to non-Euclidean norms. Studying the online and stochastic convex optimization setting (Section 3), in addition to ERM (Section 2), has several advantages. First, it allows us to obtain a learning guarantee for an efficient single-pass learning methods, namely stochastic gradient descent (or mirror descent), as well as for the non-stochastic regret. Second, the bound we obtain in the convex optimization setting (Section 3) is actually better then the bound for the ERM (Section 2) as it avoids all polylogarithmic and large constant factors. Third, the bound is applicable to other non-negative online or stochastic optimization problems beyond classification, including problems for which ERM is not applicable (see, e.g., [24]).

The detailed proofs of the statements claimed in this paper can be found in the supplementary material corresponding to the paper.

2 Empirical Risk Minimization with Smooth Loss

Recall that the Rademacher complexity of $\mathcal{H}$ for any $n \in \mathbb{N}$ given by [2]:

$$\mathcal{R}_n(\mathcal{H}) = \sup_{x_1,\ldots,x_n \in \mathcal{X}} \mathbb{E}_{\sigma \sim \text{Unif}(\{\pm 1\}^n)} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \left|\sum_{i=1}^{n} h(x_i)\sigma_i\right|\right]. \tag{4}$$

Throughout we shall consider the “worst case” Rademacher complexity.

Our starting point is the learning bound (1) that applies to D-Lipschitz loss functions, i.e. such that $|\phi'(t,y)| \leq D$ (we always take derivatives w.r.t. the first argument). What type of bound can we obtain if we instead bound the second derivative $\phi''(t,y)$? We will actually avoid talking about the second derivative explicitly, and instead say that a function is H-smooth iff its derivative is H-Lipschitz. For twice differentiable ϕ, this just means that $|\phi''| \leq H$. The central observation, which allows us to obtain guarantees for smooth loss functions, is that for a smooth loss, the derivative can be bounded in terms of the function value:

Lemma 2.1. *For an H-smooth non-negative function $f : \mathbb{R} \mapsto \mathbb{R}$, we have: $|f'(t)| \leq \sqrt{4Hf(t)}$*

This Lemma allows us to argue that close to the optimum value, where the *value* of the loss is small, then so is its derivative. Looking at the dependence of (1) on the derivative bound D, we are guided by the following heuristic intuition: Since we should be concerned only with the behavior around the ERM, perhaps it is enough to bound $\phi'(\hat{\mathbf{w}}, x)$ at the ERM $\hat{\mathbf{w}}$. Applying Lemma 2.1 to $L(\hat{h})$, we can bound $|\mathbb{E}[\phi'(\hat{\mathbf{w}}, X)]| \leq \sqrt{4HL(\hat{h})}$. What we would actually want is to bound each $|\phi'(\hat{\mathbf{w}}, x)|$ separately, or at least have the absolute value *inside* the expectation—this is where the non-negativity of the loss plays an important role. Ignoring this important issue for the moment and plugging this instead of D into (1) yields $L(\hat{h}) \leq L^* + 4\sqrt{HL(\hat{h})}\mathcal{R}_n(\mathcal{H})$. Solving for $L(\hat{h})$ yields the desired bound (3).

This rough intuition is captured by the following Theorem:

Theorem 1. *For an H-smooth non-negative loss ϕ s.t.$\forall_{x,y,h}\, |\phi(h(x),y)| \leq b$, for any $\delta > 0$ we have that with probability at least $1-\delta$ over a random sample of size n, for any $h \in \mathcal{H}$,*

$$L(h) \leq \hat{L}(h) + K\left(\sqrt{\hat{L}(h)}\left(\sqrt{H}\log^{1.5} n\, \mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{b\log(1/\delta)}{n}}\right) + H\log^3 n\, \mathcal{R}_n^2(\mathcal{H}) + \frac{b\log(1/\delta)}{n}\right)$$

and so:

$$L\left(\hat{h}\right) \leq L^* + K\left(\sqrt{L^*}\left(\sqrt{H}\log^{1.5} n\, \mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{b\log(1/\delta)}{n}}\right) + H\log^3 n\, \mathcal{R}_n^2(\mathcal{H}) + \frac{b\log(1/\delta)}{n}\right)$$

where $K < 10^5$ is a numeric constant derived from [20] and [6].

Note that only the “confidence” terms depended on $b = \sup|\phi|$, and this is typically not the dominant term—we believe it is possible to also obtain a bound that holds in expectation over the sample (rather than with high probability) and that avoids a direct dependence on $\sup|\phi|$.

To prove Theorem 1 we use the notion of Local Rademacher Complexity [3], which allows us to focus on the behavior close to the ERM. To this end, consider the following empirically restricted loss class

$$\mathcal{L}_\phi(r) := \left\{(x,y) \mapsto \phi(h(x),y) : h \in \mathcal{H}, \hat{L}(h) \leq r\right\}$$

Lemma 2.2, presented below, solidifies the heuristic intuition discussed above, by showing that the Rademacher complexity of $\mathcal{L}_\phi(r)$ scales with $\sqrt{Hr}$. The Lemma can be seen as a higher-order version of the Lipschitz Composition Lemma [2], which states that the Rademacher complexity of the *unrestricted* loss class is bounded by $D\mathcal{R}_n(\mathcal{H})$. Here, we use the second, rather then first, derivative, and obtain a bound that depends on the empirical restriction:

Lemma 2.2. *For a non-negative H-smooth loss ϕ bounded by b and any function class $\mathcal{H}$ bounded by B:*

$$\mathcal{R}_n(\mathcal{L}_\phi(r)) \leq \sqrt{12Hr}\, \mathcal{R}_n(\mathcal{H})\left(16\log^{3/2}\left(\frac{nB}{\mathcal{R}_n(\mathcal{H})}\right) - 14\log^{3/2}\left(\frac{n\sqrt{12H}B}{\sqrt{b}}\right)\right)$$

Applying Lemma 2.2, Theorem 1 follows using standard Local Rademacher argument [3].

2.1 Related Results

Rates faster than $1/\sqrt{n}$ have been previously explored under various conditions, including when L^* is small.

The Finite Dimensional Case : Lee et al [16] showed faster rates for squared loss, exploiting the strong convexity of this loss function, even when $L^* > 0$, but only with finite VC-subgraph-dimension. Panchenko [22] provides fast rate results for general Lipschitz bounded loss functions, still in the finite VC-subgraph-dimension case. Bousquet [6] provided similar guarantees for linear predictors in Hilbert spaces when the spectrum of the kernel matrix (covariance of X) is exponentially decaying, making the situation almost finite dimensional. All these methods rely on finiteness of effective dimension to provide fast rates. In this case, smoothness is not necessary. Our method, on the other hand, establishes fast rates, when $L^* = 0$, for function classes that do *not* have finite VC-subgraph-dimension. We show how in this non-parametric case, smoothness is necessary and plays an important role (see also Table 1).

Aggregation : Tsybakov [29] studied learning rates for aggregation, where a predictor is chosen from the convex hull of a finite set of base predictors. This is equivalent to an ℓ_1 constraint where each base predictor is viewed as a "feature". As with ℓ_1-based analysis, since the bounds depend only logarithmically on the number of base predictors (i.e. dimensionality), and rely on the scale of change of the loss function, they are of "scale sensitive" nature. For such an aggregate classifier, Tsybakov obtained a rate of $1/n$ when zero (or small) risk is achieve by one of the base classifiers. Using Tsybakov's result, it is not enough for zero risk to be achieved by an aggregate (i.e. bounded ell_1) classifier in order to obtain the faster rate. Tsybakov's core result is thus in a sense more similar to the finite dimensional results, since it allows for a rate of $1/n$ when zero error is achieved by a finite cardinality (and hence finite dimension) class. Tsybakov then used the approximation error of a small class of base predictors w.r.t. a large hypothesis class (i.e. a covering) to obtain learning rates for the large hypothesis class by considering aggregation within the small class. However these results only imply fast learning rates for hypothesis classes with very low complexity. Specifically to get learning rates better than $1/\sqrt{n}$ using these results, the covering number of the hypothesis class at scale ϵ needs to behave as $1/\epsilon^p$ for some $p < 2$. But typical classes, including the class of linear predictors with bounded norm, have covering numbers that scale as $1/\epsilon^2$ and so these methods do not imply fast rates for such function classes. In fact, to get rates of $1/n$ with these techniques, even when $L^* = 0$, requires covering numbers that do not increase with ϵ at all, and so actually finite VC-subgraph-dimension. Chesneau et al [10] extend Tsybakov's work also to general losses, deriving similar results for Lipschitz loss function. The same caveats hold: even when $L^* = 0$, rates faster when $1/\sqrt{n}$ require covering numbers that grow slower than $1/\epsilon^2$, and rates of $1/n$ essentially require finite VC-subgraph-dimension. Our work, on the other hand, is applicable whenever the Rademacher complexity (equivalently covering numbers) can be controlled. Although it uses some similar techniques, it is also rather different from the work of Tsybakov and Chesneau et al, in that it points out the importance of smoothness for obtaining fast rates in the non-parametric case: Chesneau et al relied only on the Lipschitz constant, which we show, in Section 4, is not enough for obtaining fast rates in the non-parametric case, even when $L^* = 0$.

Local Rademacher Complexities : Bartlett et al [3] developed a general machinery for proving possible fast rates based on local Rademacher complexities. However, it is important to note that the localized complexity term typically dominates the rate and still needs to be controlled. For example, Steinwart [27] used Local Rademacher Complexity to provide fast rate on the 0/1 loss of Support Vector Machines (SVMs) (ℓ_2-regularized hinge-loss minimization) based on the so called "geometric margin condition" and Tsybakov's margin condition. Steinwart's analysis is specific to SVMs. We also use Local Rademacher Complexities in order to obtain fast rates, but do so for general hypothesis classes, based only on the standard Rademacher complexity $\mathcal{R}_n(\mathcal{H})$ of the hypothesis classes, as well as the smoothness of the loss function and the magnitude of L^*, but without any further assumptions on the hypothesis classes itself.

Non-Lipschitz Loss : Beyond the strong connections between smoothness and fast rates which we highlight, we are also not aware of prior work providing an explicit and easy-to-use result for controlling a generic non-Lipschitz loss (such as the squared loss) solely in terms of the Rademacher complexity.

3 Online and Stochastic Optimization of Smooth Convex Objectives

We now turn to online and stochastic convex optimization. In these settings a learner chooses $\mathbf{w} \in \mathbf{W}$, where $\mathbf{W}$ is a closed convex set in a normed vector space, attempting to minimize an objective $\ell(\mathbf{w}, z)$ on instances $z \in \mathcal{Z}$, where $\ell : \mathbf{W} \times \mathcal{Z} \to \mathbb{R}$ is an objective function which is convex in $\mathbf{w}$. This captures learning linear predictors w.r.t. a convex loss function $\phi(t, z)$, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\ell(\mathbf{w}, (x, y)) = \phi(\langle \mathbf{w}, x \rangle, y)$, and extends beyond supervised learning.

We consider the case where the objective $\ell(\mathbf{w}, z)$ is H-smooth w.r.t. some norm $\|\mathbf{w}\|$ (the reader may choose to think of $\mathbf{W}$ as a subset of a Euclidean or Hilbert space, and $\|\mathbf{w}\|$ as the ℓ_2-norm): By this we mean that for any $z \in \mathcal{Z}$, and all $\mathbf{w}, \mathbf{w}' \in \mathbf{W}$

$$\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{w}', z)\|_* \leq H \|\mathbf{w} - \mathbf{w}'\|$$

where $\|\cdot\|_*$ is the dual norm. The key here is to generalize Lemma 2.1 to smoothness w.r.t. a vector $\mathbf{w}$, rather than scalar smoothness:

Lemma 3.1. *For an H-smooth non-negative $f : \mathbf{W} \to \mathbb{R}$, for all $\mathbf{w} \in \mathbf{W}$: $\|\nabla f(\mathbf{w})\|_* \leq \sqrt{4Hf(\mathbf{w})}$*

In order to consider general norms, we will also need to rely on a non-negative regularizer $F : \mathbf{W} \mapsto \mathbb{R}$ that is a 1-strongly convex (see Definition in e.g. [31]) w.r.t. to the norm $\|\mathbf{w}\|$ for all $\mathbf{w} \in \mathbf{W}$. For the Euclidean norm we can use the squared Euclidean norm regularizer: $F(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$.

3.1 Online Optimization Setting

In the online convex optimization setting we consider an n round game played between a learner and an adversary (Nature) where at each round i, the player chooses a $\mathbf{w}_i \in \mathbf{W}$ and then the adversary picks a $z_i \in \mathcal{Z}$. The player's choice $\mathbf{w}_i$ may only depend on the adversary's choices in *previous* rounds. The goal of the player is to have low average objective value $\frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}_i, z_i)$ compared to the best single choice in hind sight [9].

A classic algorithm for this setting is Mirror Descent [4], which starts at some arbitrary $\mathbf{w}_1 \in \mathbf{W}$ and updates $\mathbf{w}_{i+1}$ according to z_i and a stepsize η (to be discussed later) as follows:

$$\mathbf{w}_{i+1} \leftarrow \arg\min_{\mathbf{w} \in \mathbf{W}} \langle \eta \nabla \ell(\mathbf{w}_i, z_i) - \nabla F(\mathbf{w}_i), \mathbf{w}\rangle + F(\mathbf{w}) \tag{5}$$

For the Euclidean norm with $F(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$, the update (5) becomes projected online gradient descent [32]: $\mathbf{w}_{i+1} \leftarrow \Pi_{\mathbf{W}}(\mathbf{w}_i - \eta\nabla\ell(\mathbf{w}_i, z_i))$ where $\Pi_{\mathbf{W}}(\mathbf{w}) = \arg\min_{\mathbf{w}' \in \mathbf{W}} \|\mathbf{w} - \mathbf{w}'\|$ is the projection onto $\mathbf{W}$.

Theorem 2. *For any $B \in \mathbb{R}$ and $\overline{L^*}$ if we use stepsize $\eta = \frac{1}{HB^2 + \sqrt{H^2B^4 + HB^2 n \overline{L^*}}}$ for the Mirror Descent algorithm then for any instance sequence $z_1, \ldots, z_n \in \mathcal{Z}$, the average regret w.r.t. any $\mathbf{w}^* \in \mathbf{W}$ s.t. $F(\mathbf{w}^*) \leq B^2$ and $\frac{1}{n}\sum_{j=1}^n \ell(\mathbf{w}^*, z_i) \leq \overline{L^*}$ is bounded by:*

$$\frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}_i, z_i) - \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}^*, z_i) \leq \frac{4HB^2}{n} + 2\sqrt{\frac{HB^2\overline{L^*}}{n}}$$

Note that the stepsize depends on the bound $\overline{L^*}$ on the loss in hindsight. The above theorem can be proved using Lemma 3.1 and Theorem 1 of [26].

3.2 Stochastic Optimization

An online algorithm can also serve as an efficient one-pass learning algorithm in the stochastic setting. Here, we again consider an i.i.d. sample $z_1, \ldots, z_n$ from some unknown distribution (as in Section 2), and we would like to find $\mathbf{w}$ with low risk $L(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}, Z)]$. When $z = (\mathbf{x}, y)$ and $\ell(\mathbf{w}, z) = \phi(\langle\mathbf{w}, \mathbf{x}\rangle, y)$ this agrees with the supervised learning risk discussed in the Introduction and analyzed in Section 2. But instead of focusing on the ERM, we run Mirror Descent on the sample, and then take $\tilde{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_i$. Standard arguments [8] allow us to convert the online regret bound of Theorem 2 to a bound on the excess risk:

Corollary 3. *For any $B \in \mathbb{R}$ and $\overline{L^*}$, if we run Mirror Descent on the sample with $\eta = \frac{1}{HB^2 + \sqrt{H^2B^4 + HB^2 n \overline{L^*}}}$, then for any $\mathbf{w}^* \in \mathbf{W}$ with $F(\mathbf{w}^*) \leq B^2$ and $L(\mathbf{w}^*) \leq \overline{L^*}$, with expectation over the sample:*

$$L\left(\tilde{\mathbf{w}}_n\right) - L\left(\mathbf{w}^\star\right) \leq \frac{4HB^2}{n} + 2\sqrt{\frac{HB^2\overline{L^*}}{n}}.$$

It is instructive to contrast this guarantee with similar looking guarantees derived recently in the stochastic convex optimization literature [14]. There, the model is stochastic first-order optimization, i.e. the learner gets to see an unbiased estimate $\nabla l(\mathbf{w}, z_i)$ of the gradient of $L(\mathbf{w})$. The variance of the estimate is assumed to be bounded by σ^2. The expected accuracy after n gradient evaluations then has two terms: a "accelerated" term that is $O(H/n^2)$ and a slow $O(\sigma/\sqrt{n})$ term. While this result is applicable more generally (since it doesn't require non-negativity of ℓ), it is not immediately clear if our guarantees can be derived using it. The main difficulty is that σ depends on the norm of the gradient estimates. Thus, it cannot be bounded in advance even if we know that $L(\mathbf{w}^\star)$ is small. That said, it is intuitively clear that towards the end of the optimization process, the gradient norms will typically be small if $L(\mathbf{w}^\star)$ is small because of the self bounding property (Lemma 3.1).

It is interesting to note that using stability arguments, a guarantee very similar to Corollary 3, avoiding the polylogarithmic factors of Theorem 1 as well as the dependence on the bound on the loss, can be obtained also for a "batch" learning rule similar to ERM, but incorporating regularization. For given regularization parameter $\lambda > 0$ define the regularized empirical loss as $\hat{L}_\lambda(\mathbf{w}) := \hat{L}(\mathbf{w}) + \lambda F(\mathbf{w})$ and consider the Regularized Empirical Risk Minimizer

$$\hat{\mathbf{w}}_\lambda = \arg\min_{\mathbf{w}\in\mathbf{W}} \hat{L}_\lambda(\mathbf{w}) \tag{6}$$

The following theorem provides a bound on excess risk similar to Corollary 3:

Theorem 4. *For any $B \in \mathbb{R}$ and $\overline{L^*}$ if we set $\lambda = \frac{128H}{n} + \sqrt{\frac{128^2H^2}{n^2} + \frac{128H\overline{L^*}}{nB^2}}$ then for all $\mathbf{w}^\star \in \mathbf{W}$ with $F(\mathbf{w}^\star) \leq B^2$ and $L(\mathbf{w}^\star) \leq \overline{L^*}$, we have that in expectation over sample of size n:*

$$L\left(\hat{\mathbf{w}}_\lambda\right) - L\left(\mathbf{w}^\star\right) \leq \frac{256HB^2}{n} + \sqrt{\frac{2048HB^2\overline{L^*}}{n}}.$$

To prove Theorem 4 we use stability arguments similar to the ones used by Shalev-Shwartz et al [24], which are in turn based on Bousquet and Elisseeff [7]. However, while Shalev-Shwartz et al [24] use the notion of uniform stability, here it is necessary to look at stability in expectation to get the faster rates.

4 Tightness

In this Section we return to the learning rates for the ERM for parametric and for scale-sensitive hypothesis classes (i.e. in terms of the dimensionality and in terms of scale sensitive complexity measures), discussed in the Introduction and analyzed in Section 2. We compare the guarantees on the learning rates in different situations, identify differences between the parametric and scale-sensitive cases and between the smooth and non-smooth cases, and argue that these differences are real by showing that the corresponding guarantees are tight. Although we discuss the tightness of the learning guarantees for ERM in the stochastic setting, similar arguments can also be made for online learning.

Table 1 summarizes the bounds on the excess risk of the ERM implied by Theorem 1 as well previous bounds for Lipschitz loss on finite-dimensional [22] and scale-sensitive [2] classes, and a bound for squared-loss on finite-dimensional classes [9, Theorem 11.7] that can be generalized to any smooth strongly convex loss. We shall now show that the $1/\sqrt{n}$ dependencies in Table 1 are unavoidable. To do so, we will consider the class $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\| \leq 1\}$ of ℓ_2-bounded linear predictors (all norms in this Section are Euclidean), with different loss functions, and various specific distributions over $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\right\}$ and $Y = [0, 1]$. For the non-parametric lower-bounds, we will allow the dimensionality d to grow with the sample size n.

	Parametric	Scale-Sensitive
Loss function is:	$\dim(\mathcal{H}) \leq d\ ,\ \ \lvert h\rvert \leq 1$	$\mathcal{R}_n(\mathcal{H}) \leq \sqrt{R/n}$
D-Lipschitz	$\frac{dD}{n} + \sqrt{\frac{dDL^*}{n}}$	$\sqrt{\frac{D^2R}{n}}$
H-smooth	$\frac{dH}{n} + \sqrt{\frac{dHL^*}{n}}$	$\frac{HR}{n} + \sqrt{\frac{HRL^*}{n}}$
H-smooth and λ-strongly Convex	$\frac{H}{\lambda}\frac{dH}{n}$	$\frac{HR}{n} + \sqrt{\frac{HRL^*}{n}}$

Table 1: Bounds on the excess risk, up to polylogarithmic factors.

Infinite dimensional, Lipschitz (non-smooth), separable
Consider the absolute difference loss $\phi(h(\mathbf{x}), y) = |h(\mathbf{x}) - y|$, take $d = 2n$ and consider the following distribution: X is uniformly distributed over the d standard basis vectors $\mathbf{e}_i$ and if $X = \mathbf{e}_i$, then $Y = \frac{1}{\sqrt{n}}r_i$, where $r_1, \ldots, r_d \in \{\pm 1\}$ is an arbitrary sequence of signs unknown to the learner. Taking $\mathbf{w}^\star = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} r_i\mathbf{e}_i$, $\|\mathbf{w}^\star\| = 1$ and $L^* = L\left(\mathbf{w}^\star\right) = 0$. However any sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ reveals at most n of $2n$ signs r_i, and no information on the remaining signs. This means that for any learning algorithm, there exists a choice of r_i's such that on at least n of the remaining points not seen by the learner, he/she has to suffer a loss of at least $1/\sqrt{n}$, yielding an overall risk of at least $1/\sqrt{4n}$.

Infinite dimensional, smooth, non-separable, even if strongly convex
Consider the squared loss $\phi(h(\mathbf{x}), y) = (h(\mathbf{x}) - y)^2$ which is 2-smooth and 2-strongly convex. For any $\sigma \geq 0$ let $d = \sqrt{n}/\sigma$ and consider the following distribution: X is uniform over $\mathbf{e}_i$ as before, but this time $Y|X$ is random, with $Y|(X = \mathbf{e}_i) \sim \mathcal{N}(\frac{r_i}{2\sqrt{d}}, \sigma)$, where again r_i are pre-determined, unknown to the learner, random signs. The minimizer of the expected risk is $\mathbf{w}^\star = \sum_{i=1}^{d} \frac{r_i}{2\sqrt{d}} \mathbf{e}_i$, with $\|\mathbf{w}^\star\| = \frac{1}{2}$ and $L^* = L(\mathbf{w}^\star) = \sigma^2$. Furthermore, for any $\mathbf{w} \in \mathbf{W}$,

$$L\left(\mathbf{w}\right) - L\left(\mathbf{w}^\star\right) = \mathbb{E}\left[\langle \mathbf{w} - \mathbf{w}^\star, \mathbf{x}\rangle\right]^2 = \frac{1}{d}\sum_{i=1}^{d}(\mathbf{w}[i] - \mathbf{w}^\star[i])^2 = \frac{1}{d}\left\|\mathbf{w} - \mathbf{w}^\star\right\|^2$$

If the norm constraint becomes tight, i.e. $\|\hat{\mathbf{w}}\| = 1$, then $L(\hat{\mathbf{w}}) - L(\mathbf{w}^\star) \geq 1/(4d) = \sigma/(4\sqrt{n}) = \sqrt{L^*}/(4\sqrt{n})$. Otherwise, each coordinate is a separate mean estimation problem, with n_i samples, where n_i is the number of appearances of $\mathbf{e}_i$ in the sample. We have $\mathbb{E}\left[(\hat{\mathbf{w}}[i] - \mathbf{w}^\star[i])^2\right] = \sigma^2/n_i$ and so $L(\hat{\mathbf{w}}) - L^* = \frac{1}{d}\left\|\hat{\mathbf{w}} - \mathbf{w}^\star\right\|^2 = \frac{1}{d}\sum_{i=1}^{d}\frac{\sigma^2}{n_i} \geq \sqrt{\frac{L^*}{n}}$

Finite dimensional, smooth, not strongly convex, non-separable:
Take $d = 1$, with $X = 1$ with probability q and $X = 0$ with probability $1 - q$. Conditioned on $X = 0$ let $Y = 0$ deterministically and while conditioned on $X = 1$ let $Y = +1$ with probability $p = \frac{1}{2} + \frac{0.2}{\sqrt{qn}}$ and $Y = -1$ with probability $1 - p$. Consider the following 1-smooth loss :

$$\phi(h(\mathbf{x}), y) = \begin{cases} (h(\mathbf{x}) - y)^2 & \text{if } |h(\mathbf{x}) - y| \leq 1/2 \\ |h(\mathbf{x}) - y| - 1/4 & \text{if } |h(\mathbf{x}) - y| \geq 1/2 \end{cases}$$

First, irrespective of choice of $\mathbf{w}$, when $\mathbf{x} = 0$, we always have $h(\mathbf{x}) = 0$ and so suffer no loss. This happens with probability $1 - q$. Next observe that for $p > 0.5$, the optimal predictor is $\mathbf{w}^\star \geq 1/2$. However, for $n > 20$, with probability at least 0.25, $\sum_{i=1}^{n} y_i < 0$, and so$\hat{\mathbf{w}} \leq -1/2$. Hence, $L(\hat{\mathbf{w}}) - L^* > L(-1/2) - L(1/2) = \sqrt{0.16\, q/n}$. However for $p > 0.5$ and $n > 20$, $L^* > q/2$ and so with probability 0.25, $L(\hat{\mathbf{w}}) - L^* > \sqrt{0.32 L^*/n}$.

5 Implications

5.1 Improved Margin Bounds

"Margin bounds" provide a bound on the expected zero-one loss of a classifiers based on the margin $0/1$ error on the training sample. Koltchinskii and Panchenko [13] provides margin bounds for a generic class $\mathcal{H}$ based on the Rademacher complexity of the class. This is done by using a non-smooth Lipschitz "ramp" loss that upper bounds the zero-one loss and is upper-bounded by the margin zero-one loss. However, such an analysis unavoidably leads to a $1/\sqrt{n}$ rate even in the separable case. Following the same idea we use the following smooth "ramp":

$$\phi(t) = \begin{cases} 1 & t \leq 0 \\ \frac{1+\cos(\pi t/\gamma)}{2} & 0 < t < \gamma \\ 0 & t \geq \gamma \end{cases}$$

This loss function is $\frac{\pi^2}{4\gamma^2}$-smooth and is lower bounded by the zero-one loss and upper bounded by the γ margin loss. Using Theorem 1 we can now provide improved margin bounds for the zero-one loss of any classifier based on empirical margin error. Denote $\mathrm{err}(h) = \mathbb{E}\left[\mathbb{1}_{\{h(x) \neq y\}}\right]$ the zero-one risk and for any $\gamma > 0$ and sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathcal{X} \times \{\pm 1\}$ define the γ-margin empirical zero one loss as $\widehat{\mathrm{err}}_\gamma(h) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i h(\mathbf{x}_i) < \gamma\}}$.

Theorem 5. *For any hypothesis class $\mathcal{H}$, with $|h| \leq b$, and any $\delta > 0$, with probability at least $1 - \delta$, simultaneously for all margins $\gamma > 0$ and all $h \in \mathcal{H}$:*

$$\mathrm{err}(h) \leq \widehat{\mathrm{err}}_\gamma(h) + K\left(\sqrt{\widehat{\mathrm{err}}_\gamma(h)}\left(\frac{\log^{1.5} n}{\gamma}\mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(\log(\frac{4b}{\gamma})/\delta)}{n}}\right) + \frac{\log^3 n}{\gamma^2}\mathcal{R}_n^2(\mathcal{H}) + \frac{\log(\log(\frac{4b}{\gamma})/\delta)}{n}\right)$$

where K is a numeric constant from Theorem 1.

In particular, for appropriate numeric constant K :

$$\mathrm{err}(h) \leq 1.01\ \widehat{\mathrm{err}}_\gamma(h) + K\left(\frac{2\log^3 n}{\gamma^2}\mathcal{R}_n^2(\mathcal{H}) + \frac{2\log(\log(\frac{4b}{\gamma})/\delta)}{n}\right)$$

Improved margin bounds of the above form have been previously shown specifically for linear prediction in a Hilbert space based on the PAC Bayes theorem [19, 15]. However PAC-Bayes based results are specific to certain linear function class. Theorem 5, in contrast, is a generic concentration-based result that can be applied to any function class.

5.2 Interaction of Norm and Dimension

Consider the problem of learning a low-norm linear predictor with respect to the squared loss $\phi(t,z) = (t-z)^2$, where $\mathcal{X} \in \mathbb{R}^d$, for finite but very large d, and where the expected norm of X is low. Specifically, let X be Gaussian with $\mathbb{E}\,\|X\|^2 = B$, $Y = \langle \mathbf{w}^*, X\rangle + \mathcal{N}(0,\sigma^2)$ with $\|\mathbf{w}^*\| = 1$, and consider learning a linear predictor using ℓ_2 regularization. What determines the sample complexity? How does the error decrease as the sample size increases?

From a scale-sensitive statistical learning perspective, we expect that the sample complexity, and the decrease of the error, should depend on the norm B, especially if $d \gg B^2$. However, for any fixed d and B, even if $d \gg B^2$, asymptotically as the number of samples increase, the excess risk of norm-constrained or norm-regularized regression actually behaves as $L(\hat{\mathbf{w}}) - L^* \approx \frac{d}{n}\sigma^2$, and depends (to first order) only on the dimensionality d and not on B [17].

The asymptotic dependence on the dimensionality alone can be understood through Table 1. In this non-separable situation, parametric complexity controls can lead to a $1/n$ rate, ultimately dominating the $1/\sqrt{n}$ rate resulting from $L^* > 0$ when considering the scale-sensitive, non-parametric complexity control B. Combining Theorem 4 with the asymptotic $\frac{d}{n}\sigma^2$ behavior, and noting that at the worst case we can predict using a zero vector, yields the following overall picture on the expected excess risk of ridge regression with an optimally chosen λ:

$$L(\hat{\mathbf{w}}_\lambda) - L^* \le O\left(\min\left(B^2, B^2/n + B\sigma/\sqrt{n}, d\sigma^2/n\right)\right)$$

Roughly speaking, each term above describes the behavior in a different regime of the sample size. The first regime has excess risk of order B^2 which occurs until $n = \Theta(B^2)$. The second ("low-noise") regime is one where the excess risk is dominated by the norm and behaves as B^2/n, until $n = \Theta(B^2/\sigma^2)$ and $L(\hat{\mathbf{w}}) = \Theta(L^*)$. The third ("slow") regime, where the excess risk is controlled by the norm and the approximation error and behaves as $B\sigma/\sqrt{n}$, until $n = \Theta(d^2\sigma^2/B^2)$ and $L(\hat{\mathbf{w}}) = L^* + \Theta(B^2/d)$. The fourth ("asymptotic") regime is where excess risk behaves as d/n. This sheds further light on recent work by Liang and Srebro [18] based on exact asymptotics.

5.3 Sparse Prediction

The use of the ℓ_1 norm has become popular for learning sparse predictors in high dimensions, as in the LASSO. The LASSO estimator [28] $\hat{\mathbf{w}}$ is obtained by considering the squared loss $\phi(z,y) = (z-y)^2$ and minimizing $\hat{L}(\mathbf{w})$ subject to $\|\mathbf{w}\|_1 \le B$. Let us assume there is some (unknown) sparse reference predictor $\mathbf{w}^0$ that has low expected loss and sparsity (number of non-zeros) $\|\mathbf{w}^0\|_0 = k$, and that $\|\mathbf{x}\|_\infty \le 1, y \le 1$. In order to choose B and apply Theorem 1 in this setting, we need to bound $\|\mathbf{w}^0\|_1$. This can be done by, e.g., assuming that the features $\mathbf{x}[i]$ *in the support of* $\mathbf{w}^0$ are mutually uncorrelated. Under such an assumption, we have: $\|\mathbf{w}^0\|_1^2 \le k\mathbb{E}\langle \mathbf{w}^0, x\rangle^2 \le 2k(L(\mathbf{w}^0) + \mathbb{E}y^2) \le 4k$. Thus, Theorem 1 along with Rademacher complexity bounds from [11] gives us,

$$L(\hat{\mathbf{w}}) \le L(\mathbf{w}^0) + \tilde{O}\left(k\,\log(d)/n + \sqrt{k\,L(\mathbf{w}^0)\,\log(d)/n}\right). \tag{7}$$

It is possible to relax the no-correlation assumption to a bound on the correlations, as in mutual incoherence, or to other weaker conditions [25]. But in any case, unlike typical analysis for compressed sensing, where the goal is recovering $\mathbf{w}^0$ itself, here we are only concerned with correlations *inside the support of* $\mathbf{w}^0$. Furthermore, we do not require that the optimal predictor is sparse or that the model is well specified: only that there exists a low risk predictor using a small number of fairly uncorrelated features.

Bounds similar to (7) have been derived using specialized arguments [12, 30, 5]—here we demonstrate that bounds of these forms can be obtained under simple conditions, using the generic framework we suggest. It is also interesting to note that the methods and results of Section 3 can also be applied to this setting. We use the entropy regularizer

$$F(\mathbf{w}) = B\sum_i \mathbf{x}[i]\log\left(\frac{\mathbf{x}[i]}{1/d}\right) + \frac{B^2}{e} \tag{8}$$

which *is* non-negative and 1-strongly convex with respect to $\|\mathbf{w}\|_1$ on $\mathbf{W} = \left\{\mathbf{w} \in \mathbb{R}^d \middle| \mathbf{w}[i] \ge 0, \|\mathbf{w}\|_1 \le B\right\}$, with $F(\mathbf{w}) \le B^2(1 + \log d)$ (we consider here only non-negative weights—in order to allow $\mathbf{w}[i] < 0$ we can include also each feature's negation). Recalling that $\left\|\mathbf{w}^0\right\|_1 \le 2\sqrt{k}$ and using $B = 2\sqrt{k}$ in the entropy regularizer (8), we have from Theorem 4 we that $L(\hat{\mathbf{w}}_\lambda) \le L(\mathbf{w}^0) + O\left(k\,\log(d)/n + \sqrt{k\,L(\mathbf{w}^0)\,\log(d)/n}\right)$ where $\hat{\mathbf{w}}_\lambda$ is the regularized empirical minimizer (6) using the entropy regularizer (8) with λ as in Theorem 4. The advantage here is that using Theorem 4 instead of Theorem 1 avoids the extra logarithmic factors.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *FOCS*, 0:292–301, 1993.

[2] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

[3] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[4] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[5] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[6] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Ecole Polytechnique, 2002.

[7] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.

[8] N. Cesa-Bianchi, A. Conconi, and C.Gentile. On the generalization ability of on-line learning algorithms. In *NIPS*, pages 359–366, 2002.

[9] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[10] Christophe Chesneau and Guillaume Lecu. Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. 2006.

[11] S.M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, 2008.

[12] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst, H. Poincaré Probab. Statist.*, 45(1):7–57, 2009.

[13] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. of Stats.*, 30(1):1–50, 2002.

[14] G. Lan. *Convex Optimization Under Inexact First-order Information*. PhD thesis, Georgia Institute of Technology, 2009.

[15] J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems 15*, pages 423–430, 2003.

[16] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. *IEEE Trans. on Information Theory*, 1998.

[17] P. Liang, F. Bach, G. Bouchard, and M. I. Jordan. Asymptotically optimal regularization in smooth parametric models. In *NIPS*, 2010.

[18] P. Liang and N. Srebro. On the interaction between norm and dimensionality: Multiple regimes in learning. In *ICML*, 2010.

[19] D. A. McAllester. Simplified PAC-Bayesian margin bounds. In *COLT*, pages 203–215, 2003.

[20] Shahar Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Trans. On Information Theory*, 48(1):251–263, 2002.

[21] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Nauka Publishers, Moscow, 1978.

[22] D. Panchenko. Some extensions of an inequality of vapnik and chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002.

[23] David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[24] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[25] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity. Technical report, TTI-C, 2009. Available at ttic.uchicago.edu/$\sim$shai.

[26] S.Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, Hebrew University of Jerusalem, 2007.

[27] I. Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. *ANNALS OF STATISTICS*, 35:575, 2007.

[28] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

[29] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

[30] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[31] C. Zalinescu. *Convex analysis in general vector spaces*. World Scientific Publishing Co. Inc., River Edge, NJ, 2002.

[32] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.