Optimistic Rates for Learning with a Smooth Loss

Nathan Srebro Karthik Sridharan
Toyota Technological Institute at Chicago
6045 S Kenwood Ave., Chicago, IL 60637

Ambuj Tewari
Department of Computer Science
University of Texas at Austin

Abstract

We establish an excess risk bound of $\widetilde{O}\left(H\mathcal{R}_n^2 + \sqrt{HL^*}\mathcal{R}_n\right)$ for empirical risk minimization 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 $\widetilde{O}\left(RH/n\right)$ in the separable ($L^* = 0$) case and $\widetilde{O}\left(RH/n + \sqrt{L^* RH/n}\right)$ more generally. We also provide similar guarantees for online and stochastic convex optimization with a smooth non-negative objective.

1 Introduction

Consider empirical risk minimization for a hypothesis class $\mathcal{H} = \{h : \mathcal{X} \to \mathbb{R}\}$ with respect to some non-negative loss function $\phi(t, y)$. That is, we would like to learn a predictor h with small risk

$$L(h) = \mathbb{E}\left[\phi(h(X), Y)\right]$$

by minimizing the empirical risk

$$\hat{L}(h) = \frac{1}{n}\sum_{i=1}^{n} \phi(h(x_i), y_i)$$

given 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 [27] (also known as the pseudo-dimension). For such classes, learning guarantees can be obtained for any bounded loss function (i.e. any ϕ such that $|\phi| \leq b < \infty$) 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

$$\mathcal{H}_B = \left\{h_w : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle \mid \|\mathbf{w}\|_2 \leq B\right\},$$

guarantees can be obtained in terms of *scale-sensitive* measures of complexity such as fat-shattering dimensions [1], covering numbers [27] or Rademacher complexity [3]. The classical statistical learning theory approach for obtaining learning guarantees for such scale-sensitive classes is to rely on the Lipschitz constant

D of $\phi(t,y)$ with respect to its first argument t. If the loss is differentiable then this amounts to an upper bound on the magnitude of the first derivative with respect to t. The excess risk can then be bounded as (expectation here is over the sample):

$$\begin{aligned} \mathbb{E}\left[L\left(\hat{h}\right)\right] &\leq L^* + 2D\mathcal{R}_n(\mathcal{H}) \\ &= L^* + 2\sqrt{D^2\frac{R}{n}} \end{aligned} \tag{1}$$

where $\hat{h} = \arg\min_h \hat{L}(h)$ is the empirical risk minimizer (ERM), $L^* = \inf_h L(h)$ is the minimal possible risk in $\mathcal{H}$, and $\mathcal{R}_n(\mathcal{H})$ is the Rademacher complexity of the class $\mathcal{H}$. The Rademacher complexity typically scales as $\mathcal{R}_n(\mathcal{H}) = \sqrt{R/n}$, yielding the expression on the second line. For instance, in the case of ℓ_2-bounded linear predictors, $R = B^2 \|X\|_2^2$ where $\|X\|_2 = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_2$. The Rademacher complexity can be bounded by other scale-sensitive complexity measures, such as the fat-shattering dimensions and covering numbers, yielding similar guarantees in terms of these measures.

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 and logistic losses (popular for classification), or the absolute-value 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 for norm-bounded linear predictors $\mathcal{H}_B$, for instance, this results in a very disappointing excess risk bound of the form $O(\sqrt{B^4\|X\|_2^4/n})$. 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 the sample size n. The $1/\sqrt{n}$ dependence 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[1], 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 [26]:

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

Notice that the $1/\sqrt{n}$ term disappears in the separable case, and we get a graceful degredation between the $1/\sqrt{n}$ non-separable rate and the $1/n$ separable rate. The sample complexity (number of samples needed to guarantee that excess risk is smaller than ϵ) associated with the learning rate above is given by:

$$n = O\left(\frac{dD}{\epsilon}\left(\frac{L^*+\epsilon}{\epsilon}\right)\log\left(\frac{dD}{\epsilon}\right)\right) .$$

In the separable case, when $L^* = 0$, as well as in the non-separable case as long as we are concerned with excess error ϵ which is not much smaller then L^* (roughly speaking, an estimation error not much smaller than the optimal risk), the term $(L^*+\epsilon)/\epsilon$ can be thought of as constant, and the sample complexity scales roughly as $1/\epsilon$. Only when we seek excess error ϵ much smaller then L^*, might we get a $1/\epsilon^2$ scaling. We refer to such a rate as an "optimistic rate".[2]

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 $1/\sqrt{n}$ and not $1/n$, *even in the separable case*. However, for H-smooth non-negative loss functions, where the second derivative of $\phi(t,y)$ with respect

[1] Several binary classification losses evaluate to zero when the "margin" $yh(x)$ is sufficiently positive. For such losses, if the distribution is separable by some $h^* \in \mathcal{H}$ with enough margin, we have $L^* = 0$. This explains why we use the term "separable" to denote $L^* = 0$ for general losses.

[2] We have borrowed the term "optimistic" for the rates we provide from Dmitry Panchenko's lecture notes.

to t is bounded by H, a $1/n$ separable rate *is* possible. In Section 2 we obtain the following bound (with high probability) on the excess risk (up to logarithmic factors):

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

where again the second line corresponds to the typical scaling $\mathcal{R}_n(\mathcal{H}) = \sqrt{R/n}$. In this case, we obtain the following bound on the sample complexity required for excess error $L\left(\hat{h}\right) \leq L^* + \epsilon$:

$$n \leq O\left(\frac{R}{\epsilon}\left(\frac{L^*+\epsilon}{\epsilon}\right)\log^3(R/\epsilon)\right) .$$

In particular, for ℓ_2-norm-bounded linear predictors $\mathcal{H}_B$ with $\|X\|_2^2 \leq 1$, the excess risk is bounded by $\widetilde{O}(HB^2/n + \sqrt{HB^2L^*/n})$. 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 (finite dimensional) case, where a rate of $1/n$ is always possible for the squared loss, regardless of the value of L^* [19]. The differences between parametric and scale-sensitive classes, and between non-smooth, smooth and strongly convex (e.g. squared) loss functions are discussed in Section 3 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 4, we consider also the online setting, in addition to the stochastic setting, and present similar guarantees for online and stochastic convex optimization [37, 29]. The guarantees of Section 4 match equation (3) for the special case of a convex loss function and norm-bounded linear predictors, but Section 4 captures a more general setting of optimizing an arbitrary non-negative smooth convex objective (there is no separate discussion of a "predictor" and a scalar loss function in Section 4). Results in Section 4 are expressed in terms of properties of the norm, rather then a measure of statistical complexity like the Radamacher complexity as in (3) and Section 2. However, the online and stochastic convex optimization setting of Section 4 is also more restrictive, as we require the objective to be convex (while for the bound (3) we make no assumption about the convexity of the hypothesis class $\mathcal{H}$ nor the loss function ϕ).

Specifically, for a non-negative H-smooth *convex* objective (see exact definition in Section 4), over a domain bounded by B, we prove that the average online regret (and so also the 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 [37, 25], 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 4 also generalize to smoothness and boundedness with respect to non-Euclidean norms.

Studying the online and stochastic convex optimization setting (Section 4), in addition to ERM (Section 2), has several advantages. First, it allows us to obtain a learning guarantee for an efficient single-pass learning method, namely stochastic gradient descent (or mirror descent), as well as for non-stochastic regret. Second, the bound we obtain in the convex optimization setting (Section 4) 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., [29]).

In order to establish our main result we go back and forth between covering numbers and Radamacher complexity, and for this purpose we include in Appendix A results establishing tight relationships between the various complexity measures. These results might also be of independent interest to readers.

2 Empirical Risk Minimization with a Smooth Loss

Recall that the worst-case Rademacher complexity [3] of $\mathcal{H}$ for any $n \in \mathbb{N}$ is given by:

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

Throughout, we shall consider this "worst case" Rademacher complexity.

Our starting point is the learning bound (1) that applies to D-Lipschitz loss functions, i.e. such that $|\phi'(t,y)| \le D$ (we always take derivatives with respect to 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 $|\phi''| \le H$. 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:*

$$|f'(t)| \le \sqrt{4Hf(t)}\ .$$

Proof. For any $t < r$, there is an $s \in [t,r]$ for which $f(r) = f(t) + f'(s)(r-t)$. Now:

$$\begin{aligned} 0 \le f(r) &= f(t) + f'(t)(r-t) + (f'(s) - f'(t))(r-t) \\ &\le f(t) + f'(t)(r-t) + H\,|s-t|\,|r-t| \le f(t) + f'(t)(r-t) + H(r-t)^2 \end{aligned}$$

Setting $r = t - \frac{f'(t)}{2H}$ yields the desired bound. □

The above 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 argument : since we should be concerned only with the behavior around the ERM, perhaps it is enough to bound $\phi'(\hat{h}(x), y)$ at the ERM $\hat{h}$. Applying Lemma 2.1 to $L(\hat{h})$, we can bound $\left|\mathbb{E}\left[\phi'(\hat{h}(x),y)\right]\right| \le \sqrt{4HL(\hat{h})}$. 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 $L(\hat{h}) \le L^* + 4\sqrt{HL(\hat{h})}\mathcal{R}_n(\mathcal{H})$. Solving for $L(\hat{h})$ yields the desired bound (3).

This rough intuition is captured by the following theorem.

Theorem 1. *Let ϕ be an H-smooth non-negative loss s.t. $\forall \hat{y}, \hat{y}', y,\ |\phi(\hat{y},y) - \phi(\hat{y}',y)| \le b$. Then, for any $\delta > 0$ we have, with probability at least $1-\delta$ over a random sample of size n, for any $h \in \mathcal{H}$,*

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

and so:

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

where $K < 10^5$ is a numeric constant derived from [24] 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 $\sup|\phi|$.

The following simple corollary of the above theorem bounds the sample complexity of learning with smooth loss functions.

Corollary 2. *Assume that for any $n \geq 1$ the Rademacher complexity of function class $\mathcal{H}$ can be bounded as $\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{R}{n}}$. Then given any H-smooth non-negative loss ϕ bounded by b and any $\delta, \epsilon > 0$, the number of samples n, required to guarantee that with probability at least $1-\delta$, $L(\hat{h}) - L^* \leq \epsilon$ is bounded as*

$$n \leq O\left(\left(\frac{R\log^3(R/\epsilon) + b\log(1/\delta)}{\epsilon}\right)\left(\frac{L^*+\epsilon}{\epsilon}\right)\right) .$$

Remark. *With slight modifications in the proof, one can replace the $\log^3(R/\epsilon)$ term above with $\log^3(B/\epsilon)$ where $B := \sup_{\mathbf{x}\in\mathcal{X}, h\in\mathcal{H}} h(x)$ is the bound on functions in the hypothesis class $\mathcal{H}$.*

To prove Theorem 1, we use the notion of local Rademacher complexity [2], which allows us to focus on the behavior of $\mathcal{H}$ in the vicinity of 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, is the key to the proof of the main theorem and solidifies the heuristic intuition discussed above. It shows that the Rademacher complexity of class $\mathcal{L}_\phi(r)$ scales as $\sqrt{Hr}$. The lemma can be seen as a higher-order version of the Lipschitz composition lemma [3], 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 and any function class $\mathcal{H}$, we have:*

$$\mathcal{R}_n(\mathcal{L}_\phi(r)) \leq 21\sqrt{6Hr}\,\log^{\frac{3}{2}}(64\,n)\,\mathcal{R}_n(\mathcal{H}) .$$

Proof outline for Lemma 2.2. We delay the detailed proof of the lemma to the appendix and provide an outline of the proof here. In order to prove the lemma, we actually move from Rademacher complexity to covering numbers, use smoothness and Lemma 2.1 to obtain an r-dependent cover of the empirically restricted class, and then return to the Rademacher complexity. More specifically the proof is outlined as follows :

1. We use a modified version of Dudley's integral to bound the Rademacher complexity of the empirically restricted loss class in terms of the L_2-covering numbers of the class.
2. We use smoothness to get an r-dependent bound on the L_2-covering numbers of the empirically restricted loss class in terms of L_∞-covering numbers of the unrestricted hypothesis class.
3. We bound the L_∞-covering numbers of the unrestricted class in terms of its fat-shattering dimension, which in turn can be bounded in terms of its Rademacher complexity.

Proof of Theorem 1. By Theorem 6.1 of [6] (specifically the displayed equation prior to the last one in the proof of the theorem) we have that if ψ_n is any sub-root function that satisfies for all $r > 0$, $\mathcal{R}_n(\mathcal{L}_\phi(r)) \leq \psi_n(r)$ then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in \mathcal{H}$,

$$L(h) \leq \hat{L}(h) + 45r_n^* + \sqrt{L(h)}\left(\sqrt{8r_n^*} + \sqrt{\frac{4b(\log\left(\frac{1}{\delta}\right) + 6\log\log n)}{n}}\right) + \frac{20b(\log\left(\frac{1}{\delta}\right) + 6\log\log n)}{n} \tag{5}$$

where r_n^* is the largest solution to equation $\psi_n(r) = r$. Now by Lemma 2.2 we have that $\psi_n(r) = 21\sqrt{6Hr}\log^{1.5} n\hat{\mathcal{R}}_n(\mathcal{H})$ satisfies the property that for all $r > 0$, $\mathcal{R}_n(\mathcal{L}_\phi(r)) \le \psi_n(r)$ and so using this we see that

$$r_n^* = 2646H\log^3(64\,n)\mathcal{R}_n^2(\mathcal{H})$$

and for this r_n^*, the upper bound (5) holds. Now using the simple fact that for any non-negative A, B, C,

$$A \le B + C\sqrt{A} \Rightarrow A \le B + C^2 + \sqrt{B}C$$

we conclude,

$$L\,(h) \le \hat{L}(h) + 106\; r_n^* + \frac{48b}{n}\left(\log\tfrac{1}{\delta} + \log\log\; n\right) + \sqrt{\hat{L}(h)\left(8r_n^* + \frac{4b}{n}\left(\log\tfrac{1}{\delta} + \log\log\; n\right)\right)}\;. \tag{6}$$

Now we claim that $\frac{4b\log\log\; n}{n} \le 0.049r_n^*$. To see this first note that by definition of b,

$$b = \max_{y,\hat{y},\hat{y}'}\left(\phi(\hat{y}, y) - \phi(\hat{y}', y)\right) \le \max_{y,\hat{y},\hat{y}'}|\phi'(\hat{y}, y)|\,|\hat{y} - \hat{y}'|$$

Now notice that in the proof of Lemma 2.1 we in fact first showed that $|f'(t)| \le \sqrt{4H(f(t) - f(r))}$ for any $r > t$ and only then using the fact that f is non-negative we concluded that $|f'(t)| \le \sqrt{4Hf(t)}$. Hence we can conclude that $|\phi'(\hat{y}, y)| \le \sqrt{4Hb}$. Hence using this in the above inequality we can conclude that

$$b \le 4H\max_{\hat{y},\hat{y}'}(\hat{y} - \hat{y}')^2 \le 16H\max_{\hat{y}}|\hat{y}|^2 = 16H\max_{x,h\in\mathcal{H}}|h(x)|^2$$

Now on the other hand by definition of Rademacher complexity and by Khintchine's inequality we have that $\mathcal{R}_n(\mathcal{H}) \ge \sup_{x,y,h\in\mathcal{H}}|\phi(h(x), y)|/\sqrt{2n}$. Thus we have shown that

$$\frac{4b\log\log\; n}{n} \le \frac{64H\sup_{x,y,h\in\mathcal{H}}|\phi(h(x), y)|^2\log\log n}{n} \le 128H\log\log n\;\mathcal{R}_n^2(\mathcal{H}) \le 0.049r_n^*$$

Plugging this back in Equation 6 we see that

$$L\,(h) \le \hat{L}(h) + 109\; r_n^* + \frac{48b\log\frac{1}{\delta}}{n} + \sqrt{\hat{L}(h)\left(9r_n^* + \frac{4b\log\frac{1}{\delta}}{n}\right)}\;.$$

Plugging in the value of $r_n^* = 2646H\log^3(64\,n)\mathcal{R}_n^2(\mathcal{H})$ we get the first inequality. To get the second inequality, we simply use the first inequality with the ERM $\hat{h}$ and further note that $\hat{L}(\hat{h}) \le \hat{L}(h^*)$ (where h^* is $\operatorname{argmin}_{h\in\mathcal{H}} L(h)$). This gives us a bound of

$$L\left(\hat{h}\right) \le \hat{L}(h^*) + 109\; r_n^* + \frac{48b\log\frac{1}{\delta}}{n} + \sqrt{\hat{L}(h^*)\left(9r_n^* + \frac{4b\log\frac{1}{\delta}}{n}\right)}\;. \tag{7}$$

Now to conclude the proof notice that by Bernstein's inequality, with probability at least $1 - \delta$:

$$\begin{aligned}\hat{L}(h^*) - L^* &\le \sqrt{\frac{4\mathbb{E}\left[(\phi(h^*(x), y) - L\,(h^*))^2\right]\log\frac{1}{\delta}}{n}} + \frac{4b\log\frac{1}{\delta}}{n}\\ &\le \sqrt{\frac{8bL\,(h^*)\log\frac{1}{\delta}}{n}} + \frac{4b\log\frac{1}{\delta}}{n}\end{aligned} \tag{8}$$

Hence using the above in Equation 7 we get that

$$L\left(\hat{h}\right) \leq L^* + 109\ r_n^* + \frac{52b\log\frac{1}{\delta}}{n} + \sqrt{\hat{L}(h^*)\left(9r_n^* + \frac{4b\log\frac{1}{\delta}}{n}\right)} + \sqrt{\frac{8bL^*\log\frac{1}{\delta}}{n}}\ .$$

Again Equation 8 implies that with probability $1-\delta$, $\hat{L}(h^*) \leq \frac{3}{2}L^* + \frac{8b\log\frac{1}{\delta}}{n}$ and so using this in the above we conclude that

$$L\left(\hat{h}\right) \leq L^* + 109\ r_n^* + \frac{52b\log\frac{1}{\delta}}{n} + \sqrt{\left(\frac{3}{2}L^* + \frac{8b\log\frac{1}{\delta}}{n}\right)\left(9r_n^* + \frac{4b\log\frac{1}{\delta}}{n}\right)} + \sqrt{\frac{8bL^*\log\frac{1}{\delta}}{n}}\ .$$

Plugging in r_n^* and over bounding with appropriate numeric constant K concludes the proof. □

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 [19] 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. [26] provides optimistic rate results for general Lipschitz bounded loss functions, still in the finite VC-subgraph-dimension case. [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 optimistic rates (and a fast rate when $L^* = 0$), for function classes that do *not* have finite VC-subgraph-dimension. In Section 3 We show how in the non-parametric case, smoothness is necessary for optimistic rates and how it plays an important role (see also Table 1).

Aggregation [34] 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 a "scale sensitive" nature. For such an aggregate classifier, Tsybakov obtained a rate of $1/n$ when zero (or small) risk is achieved by one of the base classifiers. In Tsybakov's result, it is not enough to assume that zero risk is achieved by an aggregate (i.e. bounded ℓ_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 with respect to 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 requires finite VC-subgraph-dimension.

[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 3, is not enough for obtaining fast rates in the non-parametric case, even when $L^* = 0$.

Local Rademacher Complexities [2] 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, [32] 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 We are 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 A Sharp Understanding of Slow, Optimistic and Fast Rates

In this section we look at 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 for which algorithms and upper bounds are provided in the next section.

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 [26] and scale-sensitive [3] 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.

Loss function is:	Parametric $\dim(\mathcal{H}) \le d$, $\lvert h\rvert \le 1$	Scale-Sensitive $\mathcal{R}_n(\mathcal{H}) \le \sqrt{R/n}$
D-Lipschitz	$\frac{dD}{n} + \sqrt{\frac{dDL^*}{n}}$	$\sqrt{\frac{D^2 R}{n}}$
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.

We shall now show that the $1/\sqrt{n}$ dependencies in Table 1 are unavoidable. To do so, we will consider the class $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_2 \le 1\}$ 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 $\mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le 1\right\}$ and $Y = [0, 1]$. For the non-parametric lower-bounds, we will allow the dimensionality d to grow with the sample size n.

Infinite dimensional, Lipschitz (non-smooth), $\mathbf{L^* = 0}$ case
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 (say drawn randomly beforehand). Taking $\mathbf{w}^\star = \frac{1}{\sqrt{n}} \sum_{i=1}^n r_i \mathbf{e}_i$, $\|\mathbf{w}^\star\|_2 = 1$ and $L^* = L(\mathbf{w}^\star) = 0$. 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 $\geq n$ signs. This means that for any algorithm used by the learner, there exists a choice of r_i's such that on at least n of the remaining points not seen by the learner the learner has to suffer a loss of at least $1/\sqrt{n}$, yielding an overall risk of at least $1/(2\sqrt{n})$.

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 $\mathbf{w} \in \mathbf{W}$,

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

If the norm constraint becomes tight, i.e. $\|\hat{\mathbf{w}}\|_2 = 1$, 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

$$L(\hat{\mathbf{w}}) - L^* = \frac{1}{d} \|\hat{\mathbf{w}} - \mathbf{w}^\star\|_2^2 = \frac{1}{d} \sum_{i=1}^d \frac{\sigma^2}{n_i} \geq \frac{\sigma^2}{d} \frac{d^2}{\sum_i n_i} = \frac{\sigma^2 d}{n} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{L^*}{n}}$$

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 $X = 0$ let $Y = 0$ deterministically, 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 function, which is quadratic around the correct prediction, but linear away from it:

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

First note that irrespective of choice of $\mathbf{w}$, when $\mathbf{x} = 0$ and so $y = 0$ we always have $h(\mathbf{x}) = 0$ and so suffer no loss. This happens with probability $1 - q$. Next observe that for $p > 1/2$, 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 the empirical minimizer is $\hat{\mathbf{w}} \leq -1/2$. We can now calculate

$$L(\hat{\mathbf{w}}) - L^* > L(-1/2) - L(1/2) = q(2p - 1) + (1 - q)0 = \frac{0.4\, q}{\sqrt{qn}} = \frac{0.4\, \sqrt{q}}{\sqrt{n}}.$$

However note that for $p > 1/2$, $\mathbf{w}^* = \frac{3}{2} - \frac{1}{2p}$ and so for $n > 20$, $L^* > \frac{q}{2}$. Hence we conclude that with probability 0.25 over the sample,

$$L(\hat{\mathbf{w}}) - L^* > \sqrt{\frac{0.32 L^*}{n}}.$$

4 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 (loss) $\ell(\mathbf{w}, z)$ 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 using 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 well 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 an 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 $\mathbf{w}, \mathbf{w}' \in \mathbf{W}$

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

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. This is done by the next lemma.

Lemma 4.1. *For an H-smooth non-negative $f : \mathbf{W} \to \mathbb{R}$, for all $\mathbf{w} \in \mathbf{W}$:*

$$\|\nabla f(\mathbf{w})\|_* \le \sqrt{4Hf(\mathbf{w})}\ .$$

Proof. For any $\mathbf{w}_0$ such that $\|\mathbf{w} - \mathbf{w}_0\| \le 1$, let $g(t) = g(\mathbf{w}_0 + t(\mathbf{w} - \mathbf{w}_0))$. For any $t, s \in \mathbb{R}$,

$$\begin{aligned}
|g'(t) - g'(s)| &= |\langle \nabla f(\mathbf{w}_0 + t(\mathbf{w} - \mathbf{w}_0)) - \nabla f(\mathbf{w}_0 + s(\mathbf{w} - \mathbf{w}_0)), \mathbf{w} - \mathbf{w}_0\rangle| \\
&\le \|\nabla f(\mathbf{w}_0 + t(\mathbf{w} - \mathbf{w}_0)) - \nabla f(\mathbf{w}_0 + s(\mathbf{w} - \mathbf{w}_0))\|_* \ \|\mathbf{w} - \mathbf{w}_0\| \\
&\le H|t - s|\|\mathbf{w} - \mathbf{w}_0\|^2 \\
&\le H|t - s|
\end{aligned}$$

Hence g is H-smooth and so by Lemma 2.1 $|g'(t)| \le \sqrt{4Hg(t)}$. Setting $t = 1$ we have, $\langle \nabla f(\mathbf{w}), \mathbf{w} - \mathbf{w}_0\rangle \le \sqrt{4Hf(\mathbf{w})}$. Taking supremum over $\mathbf{w}_0$ such that $\|\mathbf{w}_0 - \mathbf{w}\| \le 1$ we conclude that

$$\|\nabla f(\mathbf{w})\|_* = \sup_{\mathbf{w}_0 : \|\mathbf{w} - \mathbf{w}_0\| \le 1} \langle \nabla f(\mathbf{w}), \mathbf{w} - \mathbf{w}_0\rangle \le \sqrt{4Hf(\mathbf{w})} \qquad \square$$

The above lemma effectively shows that smoothness implies the so called "self-bounding" property for the objective. This property is used by [28] to show optimistic type rates in the online setting. In the following sub-section, we use the self-bounding property implied by the above lemma along with result by [28] to obtain optimistic rates in the online setting.

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, e.g., [36]) with respect to the norm $\|\mathbf{w}\|$ over $\mathbf{W}$. For the Euclidean norm, we can use the squared Euclidean norm regularizer: $F(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|_2^2$.

4.1 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:

$$\mathbf{w}_{i+1} \leftarrow \arg\min_{\mathbf{w}\in\mathbf{W}} \langle \eta \nabla \ell(\mathbf{w}_i, z_i) - \nabla F(\mathbf{w}_i), \mathbf{w}\rangle + F(\mathbf{w}) \tag{9}$$

For the Euclidean norm with $F(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$, the update (9) becomes projected online gradient descent [37]:

$$\mathbf{w}_{i+1} \leftarrow \Pi_{\mathbf{W}}(\mathbf{w}_i - \eta \nabla \ell(\mathbf{w}_i, z_i)) \tag{10}$$

where $\Pi_{\mathbf{W}}(\mathbf{w}) = \arg\min_{\mathbf{w}'\in\mathbf{W}} \|\mathbf{w} - \mathbf{w}'\|_2$ is the Euclidean projection onto $\mathbf{W}$.

Equipped with Lemma 4.1 which implies self-bounding property and a result by [28] we have the following theorem that provides optimistic rates for the online learning of smooth objectives.

Theorem 3. *For any $B \in \mathbb{R}$ and $\overline{L^*}$ if we use stepsize $\eta = \frac{1}{HB^2+\sqrt{H^2B^4+HB^2n\overline{L^*}}}$ for the Mirror Descent algorithm then for any instance sequence $z_1, \ldots, z_n \in \mathcal{Z}$, the average regret with respect to any $\mathbf{w}^* \in \mathbf{W}$ such that $F(\mathbf{w}^*) \le B^2$ and $\frac{1}{n}\sum_{j=1}^n \ell(\mathbf{w}^*, z_i) \le \overline{L^*}$, is bounded by:*

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

Note that the stepsize depends on the bound $\overline{L^*}$ on the loss in hindsight.

Proof. The proof follows from Lemma 4.1 and Theorem 1 of [28], using $U_1 = B^2$ and $U_2 = n\overline{L^*}$ in the Theorem. □

4.2 Stochastic Optimization I: Stochastic Mirror Descent

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 (or Projected Online Gradient Descent in case of a Euclidean norm) 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 3 to a bound on the excess risk:

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

$$\mathbb{E}\left[L\left(\tilde{\mathbf{w}}\right)\right] - L\left(\mathbf{w}^\star\right) \le \frac{4HB^2}{n} + 2\sqrt{\frac{HB^2\overline{L^*}}{n}},$$

where the expectation is over the sample.

Again, one must know a bound $\overline{L^*}$ on the risk in order to choose the stepsize.

It is instructive to contrast this guarantee with similar looking guarantees derived recently in the stochastic convex optimization literature [17]. 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 does not 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 4.1). Exploring this connection can be fruitful direction for further research.

4.3 Stochastic Optimization II: Regularized Batch Optimization

It is interesting to note that using stability arguments, a guarantee very similar to Corollary 4, avoiding the polylogarithmic factors of Theorem 1 as well as the dependence on the bound on the loss (b in Theorem 1), can be obtained also for a "batch" learning rule similar to ERM, but incorporating penalty-type regularization. For a 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

$$\hat{\mathbf{w}}_\lambda = \arg\min_{\mathbf{w}\in\mathbf{W}} \hat{L}_\lambda(\mathbf{w}) \tag{11}$$

The following theorem provides a bound on excess risk similar to Corollary 4:

Theorem 5. *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 $F(\mathbf{w}^\star) \le B^2$ and $L(\mathbf{w}^\star) \le \overline{L^*}$, we have*

$$\mathbb{E}\left[L\left(\hat{\mathbf{w}}_\lambda\right)\right] - L\left(\mathbf{w}^\star\right) \le \frac{256HB^2}{n} + \sqrt{\frac{2048HB^2\overline{L^*}}{n}} ,$$

where the expectation is over the sample of size n.

To prove Theorem 5, we use stability arguments similar to the ones used by [29], which are in turn based on [7]. However, while [29] use the notion of uniform stability, here it is necessary to look at stability in expectation to get the faster rates (uniform stability does not hold with the desired rate).

To use stability based arguments, for each $i \in [n]$ we consider a perturbed sample where instance z_i is replaced by instance z_i' drawn independently from same distribution as z_i. Let $\hat{L}^{(i)}(\mathbf{w}) = \frac{1}{n}(\sum_{j\neq i} \ell(\mathbf{w}, z_j) + \ell(\mathbf{w}, z_i'))$ be the empirical risk over the perturbed sample, and consider the corresponding regularized empirical risk minimizer $\hat{\mathbf{w}}_\lambda^{(i)} = \arg\min_{\mathbf{w}} \hat{L}_\lambda^{(i)}(\mathbf{w})$, where $\hat{L}_\lambda^{(i)}(\mathbf{w}) = \hat{L}^{(i)}(\mathbf{w}) + \lambda F(\mathbf{w})$. We first prove the following lemma on the expected stability of the regularized minimizer.

Lemma 4.2. *For any $i \in [n]$ we have that*

$$\mathbb{E}_{z_1,\ldots,z_n,z_i'}\left[\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i) - \ell(\hat{\mathbf{w}}_\lambda, z_i)\right] \le \frac{32H}{\lambda n}\mathbb{E}_{z_1,\ldots,z_n}\left[L(\hat{\mathbf{w}}_\lambda)\right] .$$

Proof.

$$\begin{aligned}
\hat{L}_\lambda(\hat{\mathbf{w}}_\lambda^{(i)}) - \hat{L}_\lambda(\hat{\mathbf{w}}_\lambda) &= \frac{\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i) - \ell(\hat{\mathbf{w}}_\lambda, z_i)}{n} + \frac{\ell(\hat{\mathbf{w}}_\lambda, z_i') - \ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i')}{n} + \hat{L}_\lambda^{(i)}(\hat{\mathbf{w}}_\lambda^{(i)}) - \hat{L}_\lambda^{(i)}(\hat{\mathbf{w}}_\lambda) \\
&\le \frac{\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i) - \ell(\hat{\mathbf{w}}_\lambda, z_i)}{n} + \frac{\ell(\hat{\mathbf{w}}_\lambda, z_i') - \ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i')}{n} \\
&\le \frac{1}{n}\|\hat{\mathbf{w}}_\lambda^{(i)} - \hat{\mathbf{w}}_\lambda\| \left(\|\nabla\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i)\|_* + \|\nabla\ell(\hat{\mathbf{w}}_\lambda, z_i')\|_*\right) \\
&\le \frac{2\sqrt{H}}{n}\|\hat{\mathbf{w}}_\lambda^{(i)} - \hat{\mathbf{w}}_\lambda\| \left(\sqrt{\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i)} + \sqrt{\ell(\hat{\mathbf{w}}_\lambda, z_i')}\right)
\end{aligned}$$

where the last inequality follows from Lemma 4.1. By λ-strong convexity of $\hat{L}_\lambda$ we have that

$$\hat{L}_\lambda(\hat{\mathbf{w}}_\lambda^{(i)}) - \hat{L}_\lambda(\hat{\mathbf{w}}_\lambda) \geq \frac{\lambda}{2}\|\hat{\mathbf{w}}_\lambda^{(i)} - \hat{\mathbf{w}}_\lambda\|^2.$$

We can conclude that

$$\|\hat{\mathbf{w}}_\lambda^{(i)} - \hat{\mathbf{w}}_\lambda\| \leq \frac{4\sqrt{H}}{\lambda n}\left(\sqrt{\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i)} + \sqrt{\ell(\hat{\mathbf{w}}_\lambda, z_i')}\right)$$

This gives us:

$$\begin{aligned}
\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i) - \ell(\hat{\mathbf{w}}_\lambda, z_i) &\leq \|\nabla \ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i)\|_* \|\hat{\mathbf{w}}_\lambda^{(i)} - \hat{\mathbf{w}}_\lambda\| \\
&\leq \sqrt{4H\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i)}\left(\frac{4\sqrt{H}}{\lambda}\left(\sqrt{\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i)} + \sqrt{\ell(\hat{\mathbf{w}}_\lambda, z_i')}\right)\right) \\
&\leq \frac{16H}{\lambda n}\left(\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i) + \ell(\hat{\mathbf{w}}_\lambda, z_i')\right)
\end{aligned}$$

Taking expectation:

$$\begin{aligned}
\mathbb{E}_{z_1,\ldots,z_n,z_i'}\left[\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i) - \ell(\hat{\mathbf{w}}_\lambda, z_i)\right] &\leq \frac{16H}{\lambda n}\mathbb{E}_{z_1,\ldots,z_n,z_i'}\left[\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i) + \ell(\hat{\mathbf{w}}_\lambda, z_i')\right] \\
&= \frac{16H}{\lambda n}\mathbb{E}_{z_1,\ldots,z_n,z_i'}\left[L\left(\hat{\mathbf{w}}_\lambda^{(i)}\right) + L\left(\hat{\mathbf{w}}_\lambda\right)\right] = \frac{32H}{\lambda n}\mathbb{E}_{z_1,\ldots,z_n}\left[L\left(\hat{\mathbf{w}}_\lambda\right)\right] \quad \square
\end{aligned}$$

Proof of Theorem 5. By Lemma 4.2 we have :

$$\begin{aligned}
\mathbb{E}_{z_1,\ldots,z_n}\left[L_\lambda(\hat{\mathbf{w}}_\lambda) - L_\lambda(\mathbf{w}_\lambda^\star)\right] &\leq \mathbb{E}_{z_1,\ldots,z_n}\left[L_\lambda(\hat{\mathbf{w}}_\lambda) - \hat{L}_\lambda(\hat{\mathbf{w}}_\lambda)\right] = \mathbb{E}_{z_1,\ldots,z_n}\left[L(\hat{\mathbf{w}}_\lambda) - \hat{L}(\hat{\mathbf{w}}_\lambda)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z_1,\ldots,z_n,z_i'}\left[\ell(\hat{\mathbf{w}}_\lambda^{(i)}, z_i) - \ell(\hat{\mathbf{w}}_\lambda, z_i)\right] \leq \frac{32H}{\lambda n}\mathbb{E}_{z_1,\ldots,z_n}\left[L(\hat{\mathbf{w}}_\lambda)\right]
\end{aligned}$$

Noting the definition of $\hat{L}_\lambda(\mathbf{w})$ and rearranging we get

$$\mathbb{E}_{z_1,\ldots,z_n}\left[L(\hat{\mathbf{w}}_\lambda) - L(\mathbf{w}^\star)\right] \leq \frac{32H}{\lambda n}\mathbb{E}_{z_1,\ldots,z_n}\left[L(\hat{\mathbf{w}}_\lambda)\right] + \lambda F(\mathbf{w}^\star) - \lambda F(\hat{\mathbf{w}}_\lambda) \leq \frac{32H}{\lambda n}\mathbb{E}_{z_1,\ldots,z_n}\left[L(\hat{\mathbf{w}}_\lambda)\right] + \lambda F(\mathbf{w}^\star)$$

Rearranging further we get

$$\mathbb{E}_{z_1,\ldots,z_n}\left[L\left(\hat{\mathbf{w}}_\lambda\right)\right] - L\left(\mathbf{w}^\star\right) \leq \left(\frac{1}{1 - \frac{32H}{\lambda n}} - 1\right)L\left(\mathbf{w}^\star\right) + \frac{\lambda}{1 - \frac{32H}{\lambda n}}F(\mathbf{w}^\star)$$

plugging in the value of λ gives the result. $\square$

5 Implications

We demonstrate the implications of our results in several settings.

5.1 Improved Margin Bounds

"Margin bounds" provide a bound on the expected zero-one loss of a classifiers based on the margin zero-one error on the training sample. [15] 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, since as we discuss in Section 3, it is not possible to get a faster rate for a non-smooth loss. Following the same idea we use the following smooth "ramp":

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

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. Let

$$\mathrm{err}(h) = \mathbb{E}\left[\,\mathbb{1}_{\{h(x)\neq y\}}\right]$$

be 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 risk as

$$\widehat{\mathrm{err}}_\gamma(h) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i h(\mathbf{x}_i) < \gamma\}} .$$

Theorem 6. *For any hypothesis class $\mathcal{H}$, with $|h| \le b$, and any $\delta > 0$, with probability at least $1-\delta$, simultaneously for all margins $\gamma > 0$ and all $h \in \mathcal{H}$:*

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

where K is a numeric constant from Theorem 1

In particular, the above bound implies:

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

where K is an appropriate numeric constant.

Improved margin bounds of the above form have been previously shown specifically for linear prediction in a Hilbert space (as in Support Vector Machines) based on the PAC Bayes theorem [23, 18]. However these PAC-Bayes based results are specific to the linear function class. Theorem 6 is, in contrast, a generic concentration-based result that can be applied to any function class with and yields rates dominated by $\mathcal{R}^2(\mathcal{H})$.

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 $\mathbb{E}\left[\|X\|^2\right] = B$, $Y = \langle \mathbf{w}^*, X\rangle + \mathcal{N}(0,\sigma^2)$ 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 at all on B [20]. How does the scale sensitive complexity come into play?

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. (The dimension-dependent behavior here is actually a bit better then in the generic situation—the well-posed Gaussian model allows the bound to depend on $\sigma^2 = L^*$ rather then on $\sup(\langle \mathbf{w}, \mathbf{x}\rangle - y)^2 \approx B^2 + \sigma^2$).

Combining Theorem 5 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:

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

Roughly speaking, each term above describes the behavior in a different regime of the sample size:

- The first ("random") regime until $n = \Theta(B^2)$ where the excess risk is B^2.
- The second ("low-noise") regime, 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, where the excess risk is dominated by the dimensionality and behaves as d/n.

This sheds further light on recent work on this phenomena by Liang and Srebro based on exact asymptotics of simplified situations [21].

5.3 Sparse Prediction

The use of the ℓ_1 norm has become very popular for learning sparse predictors in high dimensions, as in the LASSO. The LASSO estimator [33] $\hat{\mathbf{w}}$ is obtained by considering the squared loss $\phi(z, y) = (z - y)^2$ and minimizing $\hat{L}(\mathbf{w})$ subject to $\|\mathbf{w}\|_1 \leq B$. 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 $\|\mathbf{x}\|_\infty \leq 1, y \leq 1$. 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: $\|\mathbf{w}^0\|_1^2 \leq k\mathbb{E}\left[\langle \mathbf{w}^0, x\rangle^2\right] \leq 2k(L(\mathbf{w}^0) + \mathbb{E}\left[y^2\right]) \leq 4k$. Thus, Theorem 1 along with Rademacher complexity bounds from [13] gives us,

$$L(\hat{\mathbf{w}}) \leq L(\mathbf{w}^0) + \widetilde{O}\left(\frac{k\ \log(d)}{n} + \sqrt{\frac{k\, L(\mathbf{w}^0)\ \log(d)}{n}}\right). \tag{12}$$

It is possible to relax the no-correlation assumption to a bound on the correlations, as in mutual incoherence, or to other weaker conditions [30]. 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 need to require that the optimal predictor is sparse or close to being sparse, or that the model is well specified: only that there exists a good (low risk) predictor using a small number of fairly uncorrelated features.

Bounds similar to (12) have been derived using specialized arguments [14, 35, 5]—here we demonstrate that a simple form of these bounds can be obtained under very simple conditions, using the generic framework we suggest.

It is also interesting to note that the methods and results of Section 4 can also be applied to this setting. But since $\|\mathbf{w}\|_1^2$ is *not* strongly convex with respect to $\|\mathbf{w}\|_1$, we must instead use the entropy regularizer

$$F(\mathbf{w}) = B \sum_i \mathbf{x}[i] \log\left(\frac{\mathbf{x}[i]}{1/d}\right) + \frac{B^2}{e} \tag{13}$$

which is 1-strongly convex w.r.t. $\|\cdot\|_1$ on $\mathbf{W} = \left\{\mathbf{w} \in \mathbb{R}^d \middle| \mathbf{w}[i] \geq 0, \|\mathbf{w}\|_1 \leq B\right\}$, with $F(\mathbf{w}) \leq B^2(1 + \log d)$ (we consider here only non-negative weights—in order to allow $\mathbf{w}[i] < 0$ we can include also each features negation, doubling the dimensionality). Recalling that $\left\|\mathbf{w}^0\right\|_1 \leq 2\sqrt{k}$ and using $B = 2\sqrt{k}$ in (13), we have from Theorem 5 we that:

$$L(\hat{\mathbf{w}}_\lambda) \leq L(\mathbf{w}^0) + O\left(\frac{k\ \log(d)}{n} + \sqrt{\frac{k\, L(\mathbf{w}^0)\ \log(d)}{n}}\right). \tag{14}$$

where $\hat{\mathbf{w}}_\lambda$ is the regularized empirical minimizer (11) using the entropy regularizer (13) with λ as in Theorem 5. The advantage here is that using Theorem 5 instead of Theorem 1 avoids the extra logarithmic factors (yielding a clean big-O dependence in (14) as opposed to big-$\widetilde{O}$ in (12)).

More interestingly, following Corollary 4, one can use stochastic mirror descent, taking steps of the form (9) with the entropy regularizer (13), to obtain the same performance guarantee as inn (14). This provides an efficient, single-pass optimization approach to sparse prediction as an alternative to batch optimization with an ℓ_1-norm constraint, and yielding the same (if not somewhat better) guarantees.

6 Discussion

We use the term "optimistic rates" as opposed to "fast rates" to distinguish between the rates of the form we get in equation (3) from the ones where $L(\hat{h}) - L^*$ is bounded only by $O(HR/n)$. Of course when L^* is smaller than $\frac{R}{n}$ then one can obtain a bound of $O(HR/n)$ for $L(\hat{h}) - L^*$ using the optimistic rates. However, in general, for optimistic rates one has an extra $\sqrt{L^* HR/n}$ term in the rate as compared to fast rates. While there is this crucial distinction between "optimistic" and "fast" rates, we would like to point out that the bound 3 can be re-written for any $a > 0$ as,

$$L(\hat{h}) \leq (1 + a)L^* + \widetilde{O}\left(\left(1 + \frac{1}{a}\right)\frac{HR}{n}\right)$$

As an example taking $a = 0.01$ this implies that $L(\hat{h}) - 1.01\, L^*$ converges as HR/n. Hence in practice especially since one tries to pick $\mathcal{H}$ so that L^* is small, the optimistic bounds implies fast learning rates.

The notion of Rademacher complexity used throughout this work is that of worst-case Rademacher complexity, that is supremum over sample of size n. With a Lipschitz loss, it is possible to obtain guarantees similar to (1) also in terms of the expected Rademacher complexity (taking an expectation over samples of size n), or even the empirical Rademacher complexity, calculated only on the specific sample observed [3]. A natural question is whether the worst case Rademacher complexity used in Theorem 1 can be replaced by the expected Rademacher complexity. The difference between worst case and expected Rademacher complexities might be crucial in certain applications. For example, [12] use our Theorem 1 to obtain guarantees on matrix completion with max-norm regularization under any arbitrary distribution. While this approach gave meaningful rates for matrix completion with max-norm, the Rademacher complexity of a trace-norm constrained class can only be meaningfully bounded *on average*.

Unfortunately, such a generalization is *not* possible: as [12] show, it is not possible to meaningfully generalize with respect to the squared loss by constraining the trace-norm, even with a uniform distribution where the expected Rademacher complexity is nicely behaved. This shows that our Theorem 1 *cannot* be restated in terms of the expected or empirical Rademacher complexity, in sharp contrast to the case of Lipschitz bounded loss. An interesting question is what happens when the loss function is Lipschitz *and* smooth (e.g. the logistic loss or smoothed hinge loss). Of course, in such cases a guarantee of the form (1) can be obtained in terms of the expected Rademacher complexity, replying only on the Lipschitz constant of the loss function. But we suspect that if the loss is Lipschitz and smooth (or bounded and smooth), it is also possible to obtain an optimistic rate similar to (3) in terms of the expected or empirical Rademacher complexity.

References

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

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

[3] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[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] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[8] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, September 2004.

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

[10] C. Chesneau and G. Lecué. Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. *Statistica Sinica*, 19:1407–1417, 2009.

[11] R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6):899–929, 1978.

[12] R. Foygel and N. Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *Proceedings of the 24th Annual Conference on Computational Learning Theory*, 2011.

[13] S.M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800, 2009.

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

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

[16] Vladimir Koltchinskii and Dmitry Panchenko. Complexities of convex combinations and bounding the generalization error in classification. *ANNALS OF STATISTICS*, 33:1455, 2005.

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

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

[19] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974 – 1980, 1998.

[20] P. Liang, F. Bach, G. Bouchard, and M. I. Jordan. Asymptotically optimal regularization in smooth parametric models. In *Advances in Neural Information Processing Systems 22*, pages 1132–1140, 2010.

[21] P. Liang and N. Srebro. On the interaction between norm and dimensionality: Multiple regimes in learning. In *Proceedings of the 27th International Conference on Machine Learning*, pages 647–654, 2010.

[22] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX(2):245–303, 2000.

[23] D. A. McAllester. Simplified PAC-Bayesian margin bounds. In *Proceedings of the 16th Annual Conference on Computational Learning Theory*, pages 203–215, 2003.

[24] Shahar Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions On Information Theory*, 48(1):251–263, 2002.

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

[26] D. Panchenko. Some extensions of an inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002.

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

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

[29] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[30] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2009.

[31] N. Srebro and K. Sridharan. Note on refined Dudley integral covering number bound, 2010. available at `http://ttic.uchicago.edu/~karthik/dudley.pdf`.

[32] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

[33] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

[34] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

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

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

[37] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

A Relating Covering Numbers, Fat Shattering Dimension, and Rademacher Complexity

Recall that the proof of the main lemma (Lemma 2.2) relies on moving between various complexity measures. To this end, we state and prove bounds on the relationship between these complexity measures, namely covering numbers, fat-shattering dimensions and the Rademacher complexity, some of which might be of independent interest. These bounds extend and refine previously existing results, but for completeness we provide full proofs for all the bounds used. Before we proceed, recall the following definitions of covering numbers and fat shattering dimension. For any $\epsilon > 0$ and function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$:

The L_2 covering number $\mathcal{N}_2(\mathcal{F}, \epsilon, n)$ is the supremum over samples $z_1, \ldots, z_n$ of the size of a minimal cover $\mathcal{C}_\epsilon$ such that $\forall f \in \mathcal{F}$, $\exists f_\epsilon \in \mathcal{C}_\epsilon$ s.t. $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(z_i) - f_\epsilon(z_i))^2} \leq \epsilon$.

The L_∞ covering number $\mathcal{N}_\infty(\mathcal{F}, \epsilon, n)$ is the supremum over samples $z_1, \ldots, z_n$ of the size of a minimal cover $\mathcal{C}_\epsilon$ such that $\forall f \in \mathcal{F}$, $\exists f_\epsilon \in \mathcal{C}_\epsilon$ s.t. $\max_{i\in[n]} |f(z_i) - f_\epsilon(z_i)| \leq \epsilon$.

The fat-shattering dimension $\mathrm{fat}_\epsilon(\mathcal{F})$ at scale ϵ is the maximum number of points ϵ-shattered by $\mathcal{F}$ (see e.g. [24]), that is largest $d \in \mathbb{N}$ such that there exists d points, $x_1, \ldots, x_d \in \mathcal{X}$ and witnesses $s_1, \ldots, s_d \in \mathbb{R}$ such that,

$$\forall \sigma_1, \ldots, \sigma_d \in \{\pm 1\}, \exists f \in F \text{ s.t. } \forall i \in [d],\ \sigma_i(f(x_i) - s_i) \geq \epsilon/2$$

We present bounds on the Rademacher complexity in terms of the L_2 covering numbers (Lemma A.1), on the L_∞ covering numbers in terms of the fat shattering dimension (Lemma A.2), and then on the fat-shattering dimension back in terms of the worst-case Rademacher complexity (Lemma A.3).

A.1 The Refined Dudley Integral: Bounding Rademacher Complexity with L_2 Covering Numbers

We shall find it simpler here to use the empirical Rademacher complexity for a given sample $x_1, \ldots, x_n$ [3]:

$$\hat{R}_n(\mathcal{H}) = \mathbb{E}_{\sigma\sim\mathrm{Unif}(\{\pm1\}^n)}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^{n} h(x_i)\sigma_i\right|\right] \tag{15}$$

and the L_2 covering number at scale $\epsilon > 0$ specific to a sample $x_1, \ldots, x_n$, denoted by $N_2(\epsilon, \mathcal{F}, (x_1, \ldots, x_n))$ as the size of a minimal cover $\mathcal{C}_\epsilon$ such that

$$\forall f \in \mathcal{F}, \exists f_\epsilon \in \mathcal{C}_\epsilon \text{ s.t. } \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(z_i) - f_\epsilon(z_i))^2} \leq \epsilon\ .$$

We will also denote $\hat{\mathbb{E}}\left[f^2\right] = \frac{1}{n}\sum_{i=1}^{n} f^2(x_i)$.

We state our bound in terms of the empirical Rademacher complexity and covering numbers. Taking a supremum over samples of size n, we get the same relationship between the worst-case Rademacher complexity and covering numbers, as is used in Section 2.

The below lemma relating Empirical Rademacher complexity and covering numbers is based on refinements of the well-known Dudley Integral [11]. The refinements provided in the below lemma use ideas from [16] and from [24].

Lemma A.1. *For any function class $\mathcal{F}$ containing functions $f : \mathcal{X} \mapsto \mathbb{R}$, we have that*

$$\hat{R}_n(\mathcal{F}) \leq \inf_{\alpha \geq 0}\left\{4\alpha + 10\int_{\alpha}^{\sup_{f\in F}\sqrt{\hat{\mathbb{E}}[f^2]}} \sqrt{\frac{\log \mathcal{N}_2(\epsilon, \mathcal{F}, (x_1, \ldots, x_n))}{n}}\,d\epsilon\right\}\ .$$

Proof. Let $\beta_0 = \sup_{f \in F} \sqrt{\hat{\mathbb{E}}\left[f^2\right]}$ and for any $j \in \mathbb{Z}_+$ let $\beta_j = 2^{-j} \sup_{f \in F} \sqrt{\hat{\mathbb{E}}\left[f^2\right]}$. The basic trick here is the idea of chaining. For each j let T_i be a (proper) L_2-cover at scale β_j of $\mathcal{F}$ for the given sample. For each $f \in \mathcal{F}$ and j, pick an $\hat{f}_i \in T_i$ such that $\hat{f}_i$ is an β_i approximation of f. Now for any N, we express f by chaining as

$$f = f - \hat{f}_N + \sum_{i=1}^{N} \left(\hat{f}_i - \hat{f}_{i-1}\right)$$

where $\hat{f}_0 = 0$. Hence for any N we have that

$$\begin{aligned}
\hat{R}_n(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(f(\mathbf{x}_i) - \hat{f}_N(\mathbf{x}_i) + \sum_{j=1}^{N}\left(\hat{f}_j(\mathbf{x}_i) - \hat{f}_{j-1}(\mathbf{x}_i)\right)\right)\right] \\
&\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(f(\mathbf{x}_i) - \hat{f}_N(\mathbf{x}_i)\right)\right] + \sum_{j=1}^{N} \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(\hat{f}_j(\mathbf{x}_i) - \hat{f}_{j-1}(\mathbf{x}_i)\right)\right] \\
&\leq \frac{1}{n}\sqrt{\sum_{i=1}^{n} \sigma_i^2}\ \sup_{f \in \mathcal{F}} \sqrt{\sum_{i=1}^{n} (f(x_i) - \hat{f}_N(x_i)^2} + \sum_{j=1}^{N} \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(\hat{f}_j(\mathbf{x}_i) - \hat{f}_{j-1}(\mathbf{x}_i)\right)\right] \\
&\leq \beta_N + \sum_{j=1}^{N} \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(\hat{f}_j(\mathbf{x}_i) - \hat{f}_{j-1}(\mathbf{x}_i)\right)\right] \qquad (16)
\end{aligned}$$

where the step before last is due to Cauchy-Shwarz inequality and $\sigma = [\sigma_1, ..., \sigma_n]^\top$. Now note that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} (\hat{f}_j(x_i) - \hat{f}_{j-1}(x_i))^2 &= \frac{1}{n}\sum_{i=1}^{n} \left((\hat{f}_j(x_i)) - f(x_i)) + (f(x_i) - \hat{f}_{j-1}(x_i))\right)^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n} \left(\hat{f}_j(x_i)) - f(x_i)\right)^2 + \frac{2}{n}\sum_{i=1}^{n} \left(f(x_i) - \hat{f}_{j-1}(x_i)\right)^2 \\
&\leq 2\beta_j^2 + 2\beta_{j-1}^2 = 6\beta_j^2 \ .
\end{aligned}$$

Now Massart's finite class lemma [22] states that if for any function class $\mathcal{G}$, $\sup_{g \in \mathcal{G}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} g(x_i)^2} \leq R$, then $\hat{R}_n(\mathcal{G}) \leq \sqrt{\frac{2R^2 \log(|\mathcal{G}|)}{n}}$. Applying this to function classes $\{f - f' : f \in T_j,\ f' \in T_{j-1}\}$ (for each j) we get from (16) that for any N,

$$\begin{aligned}
\hat{R}_n(\mathcal{F}) &\leq \beta_N + \sum_{j=1}^{N} \beta_j \sqrt{\frac{12\log(|T_j|\ |T_{j-1}|)}{n}} \\
&\leq \beta_N + \sum_{j=1}^{N} \beta_j \sqrt{\frac{24\ \log\ |T_j|}{n}} \\
&\leq \beta_N + 10\sum_{j=1}^{N} (\beta_j - \beta_{j+1}) \sqrt{\frac{\log\ |T_j|}{n}} \\
&\leq \beta_N + 10\sum_{j=1}^{N} (\beta_j - \beta_{j+1}) \sqrt{\frac{\log\ \mathcal{N}_2\left(\beta_j, \mathcal{F}, (x_1, \ldots, x_n)\right)}{n}} \\
&\leq \beta_N + 10\int_{\beta_{N+1}}^{\beta_0} \sqrt{\frac{\log\ \mathcal{N}_2\left(\epsilon, \mathcal{F}, (x_1, \ldots, x_n)\right)}{n}} d\epsilon
\end{aligned}$$

where the third step is because $2(\beta_j - \beta_{j+1}) = \beta_j$ and we bounded $\sqrt{24}$ by 5. Now for any $\alpha > 0$, pick $N = \sup\{j : \beta_j > 2\alpha\}$. In this case we see that by our choice of N, $\beta_{N+1} \leq 2\alpha$ and so $\beta_N = 2\beta_{N+1} \leq 4\epsilon$. Also note that since $\beta_N > 2\alpha$, $\beta_{N+1} = \frac{\beta_N}{2} > \alpha$. Hence we conclude that

$$\hat{R}_n(\mathcal{F}) \leq 4\alpha + 10 \int_{\alpha}^{\sup_{f\in F}\sqrt{\hat{\mathbb{E}}[f^2]}} \sqrt{\frac{\log\ \mathcal{N}_2\left(\epsilon, \mathcal{F}, (x_1, \ldots, x_n)\right)}{n}} d\epsilon\ .$$

Since the choice of α was arbitrary we take an infimum over α. □

A.2 Bounding L_∞ covering number by Fat-shattering Dimension

The following proposition and lemma are standard in statistical learning theory and their proof can be found, for instance, in [1]. We provide the statement and the proof of the lemma for completeness and so that we can state it in the exact form it is used in this work.

Proposition 7. *Let $\mathcal{H} \subseteq \{0, \ldots, k\}^{\mathcal{X}}$ be a class of functions with $\mathrm{fat}_2 = d$. Then, we have,*

$$\mathcal{N}_\infty(1/2, \mathcal{H}, n) \leq \sum_{i=0}^{d} \binom{n}{i} k^i$$

and specifically for $n \geq d$ this gives,

$$\mathcal{N}_\infty(1/2, \mathcal{H}, n) \leq \left(\frac{ekn}{d}\right)^d\ .$$

Lemma A.2. *For any function class $\mathcal{H}$ bounded by B and any $\alpha > 0$ such that $\mathrm{fat}_\alpha < n$, we have,*

$$\mathcal{N}_\infty(\alpha, \mathcal{H}, n) \leq \left(\frac{2eBn}{\alpha\ \mathrm{fat}_\alpha(\mathcal{H})}\right)^{\mathrm{fat}_\alpha(\mathcal{H})}\ .$$

Proof. For any $\alpha > 0$, define an α-discretization of the $[-B, B]$ interval as $B_\alpha = \{-B + \alpha/2, -B + 3\alpha/2, \ldots, -B + (2k+1)\alpha/2, \ldots\}$ for $0 \leq k$ and $(2k+1)\alpha \leq 4B$. Also for any $a \in [-B, B]$, define $\lfloor a \rfloor_\alpha = \operatorname*{argmin}_{r\in B_\alpha} |r - a|$ with ties being broken by choosing the smaller discretization point. For a function $h : \mathcal{X} \mapsto [-B, B]$ let the function $\lfloor h \rfloor_\alpha$ be defined pointwise as $\lfloor h(x) \rfloor_\alpha$, and let $\lfloor \mathcal{H} \rfloor_\alpha = \{\lfloor h \rfloor_\alpha : h \in \mathcal{H}\}$. First, we prove that $\mathcal{N}_\infty(\alpha, \mathcal{H}, \{x_i\}_{i=1}^n) \leq \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{H} \rfloor_\alpha, \{x_i\}_{i=1}^n)$. Indeed, suppose the set V is a minimal $\alpha/2$-cover of $\lfloor \mathcal{H} \rfloor_\alpha$ on $\{x_i\}_{i=1}^n$. That is,

$$\forall h_\alpha \in \lfloor \mathcal{H} \rfloor_\alpha,\ \exists \mathbf{v} \in V \text{ s.t.} \quad |v_i - h_\alpha(x_i)| \leq \alpha/2\ .$$

Pick any $h \in \mathcal{H}$ and let $h_\alpha = \lfloor h \rfloor_\alpha$. Then $\|h - h_\alpha\|_\infty \leq \alpha/2$ and for any $i \in [n]$

$$|h(x_i) - v_i| \leq |h(x_i) - h_\alpha(x_i)| + |h_\alpha(x_i) - v_i| \leq \alpha,$$

and so V also provides an L_∞ cover at scale α.

We conclude that $\mathcal{N}_\infty(\alpha, \mathcal{H}, \{x_i\}_{i=1}^n) \leq \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{H} \rfloor_\alpha, \{x_i\}_{i=1}^n) = \mathcal{N}_\infty(1/2, \mathcal{G}, \{x_i\}_{i=1}^n)$ where $\mathcal{G} = \frac{1}{\alpha}\lfloor \mathcal{H} \rfloor_\alpha$. The functions of $\mathcal{G}$ take on a discrete set of at most $\lfloor 2B/\alpha \rfloor + 1$ values. Obviously, by adding a constant to all the functions in $\mathcal{G}$, we can make the set of values to be $\{0, \ldots, \lfloor 2B/\alpha \rfloor\}$. We now apply Proposition 7 with an upper bound $\sum_{i=0}^{d} \binom{n}{i} k^i \leq \left(\frac{ekn}{d}\right)^d$ which holds for any $n > d$. This yields $\mathcal{N}_\infty(1/2, \mathcal{G}, \{x_i\}_{i=1}^n) \leq \left(\frac{2eBn}{\alpha \mathrm{fat}_2(\mathcal{G})}\right)^{\mathrm{fat}_2(\mathcal{G})}$.

It remains to prove $\text{fat}_2(\mathcal{G}) \leq \text{fat}_\alpha(\mathcal{H})$, or, equivalently (by scaling) $\text{fat}_{2\alpha}(\lfloor \mathcal{H} \rfloor_\alpha) \leq \text{fat}_\alpha(\mathcal{H})$. To this end, suppose there exists a set $\{x_{i=1}^n\}$ of size $d = \text{fat}_{2\alpha}(\lfloor \mathcal{H} \rfloor_\alpha)$ such that there is an witness $s_1, \ldots, s_n$ with

$$\forall \epsilon \in \{\pm 1\}^d, \ \exists h_\alpha \in \lfloor \mathcal{H} \rfloor_\alpha \quad \text{s.t. } \forall i \in [d], \ \epsilon_i (h_\alpha(x_i) - s_i) \geq \alpha \ .$$

Using the fact that for any $h \in \mathcal{H}$ and $h_\alpha = \lfloor h \rfloor_\alpha$ we have $\|h - h_\alpha\|_\infty \leq \alpha/2$, it follows that

$$\forall \epsilon \in \{\pm 1\}^d, \ \exists h \in \mathcal{H} \quad \text{s.t. } \forall i \in [d], \ \epsilon_i (h(x_i) - s_i) \geq \alpha/2 \ .$$

That is, $s_1, \ldots, s_n$ is a witness to α-shattering by $\mathcal{H}$. Thus for any $\{x_i\}_{i=1}^n$, as long as $n > \text{fat}_\alpha$

$$\mathcal{N}_\infty(\alpha, \mathcal{H}, \{x_i\}_{i=1}^n) \leq \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{H} \rfloor_\alpha, \{x_i\}_{i=1}^n) \leq \left(\frac{2eBn}{\alpha \text{fat}_{2\alpha}(\lfloor \mathcal{H} \rfloor_\alpha)} \right)^{\text{fat}_{2\alpha}(\lfloor \mathcal{H} \rfloor_\alpha)} \leq \left(\frac{2eBn}{\alpha \text{fat}_\alpha} \right)^{\text{fat}_\alpha(\mathcal{H})} \ .$$

□

A.3 Relating Fat-shattering Dimension and Rademacher complexity

The following lemma upper bounds the fat-shattering dimension at scale $\epsilon \geq \mathcal{R}_n(\mathcal{H})$ in terms of the Rademacher complexity of the function class. The proof closely follows the arguments of Mendelson [24, discussion after Definition 4.2].

Lemma A.3. *For any hypothesis class $\mathcal{H}$, any sample size n and any $\epsilon > \mathcal{R}_n(\mathcal{H})$ we have that*

$$\text{fat}_\epsilon(\mathcal{H}) \leq \frac{4 \ n \ \mathcal{R}_n(\mathcal{H})^2}{\epsilon^2} \ .$$

In particular, if $\mathcal{R}_n(\mathcal{H}) = \sqrt{R/n}$ (the typical case), then $\text{fat}_\epsilon(\mathcal{H}) \leq 4R/\epsilon^2$.

Proof. Consider any $\epsilon \geq \mathcal{R}_n(\mathcal{H})$. Let $x_1^*, \ldots, x_{\text{fat}_\epsilon}^*$ be the set of fat_ϵ shattered points. This means that there exists $s_1, \ldots, s_{\text{fat}_\epsilon}$ such that for any $J \subset [\text{fat}_\epsilon]$ there exists $h_J \in \mathcal{H}$ such that $\forall i \in J, h_J(x_i) \geq s_i + \epsilon$ and $\forall i \notin J, h_J(x_i) \leq s_i - \epsilon$. Now consider a sample $x_1, \ldots, x_{n'}$ of size $n' = \lceil \frac{n}{\text{fat}_\epsilon} \rceil \text{fat}_\epsilon$, obtained by taking each x_i^* and repeating it $\lceil \frac{n}{\text{fat}_\epsilon} \rceil$ times, i.e. $x_i = x^*_{\lfloor \frac{i}{\text{fat}_\epsilon} \rfloor}$. Now, following Mendelson's arguments:

$$\begin{aligned}
\mathcal{R}_{n'}(\mathcal{H}) &\geq \mathbb{E}_{\sigma \sim \text{Unif}\{\pm 1\}^{n'}} \left[\frac{1}{n'} \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{n'} \sigma_i h(x_i) \right| \right] \\
&\geq \frac{1}{2} \mathbb{E}_{\sigma \sim \text{Unif}\{\pm 1\}^{n'}} \left[\frac{1}{n'} \sup_{h, h' \in \mathcal{H}} \left| \sum_{i=1}^{n'} \sigma_i (h(x_i) - h'(x_i)) \right| \right] \qquad \text{(triangle inequality)} \\
&= \frac{1}{2} \mathbb{E}_{\sigma \sim \text{Unif}\{\pm 1\}^{n'}} \left[\frac{1}{n'} \sup_{h, h' \in \mathcal{H}} \left| \sum_{i=1}^{\text{fat}_\epsilon} \left(\sum_{j=1}^{\lceil n/\text{fat}_\epsilon \rceil} \sigma_{(i-1)\text{fat}_\epsilon + j} \right) (h(x_i^*) - h'(x_i^*)) \right| \right] \\
&\geq \frac{1}{2} \mathbb{E}_{\sigma \sim \text{Unif}\{\pm 1\}^{n'}} \left[\frac{1}{n'} \left| \sum_{i=1}^{\text{fat}_\epsilon} \left(\sum_{j=1}^{\lceil n/\text{fat}_\epsilon \rceil} \sigma_{(i-1)\text{fat}_\epsilon + j} \right) (h_R(x_i^*) - h_{\overline{R}}(x_i^*)) \right| \right]
\end{aligned}$$

where for each $\sigma_1, \ldots, \sigma_{n'}$, $R \subseteq [\mathrm{fat}_\epsilon]$ is given by $R = \left\{ i \middle| \mathrm{sign}\left(\sum_{j=1}^{\lceil n/\mathrm{fat}_\epsilon \rceil} \sigma_{(i-1)\lceil n/\mathrm{fat}_\epsilon \rceil + j} \right) \geq 0 \right\}$, h_R is the function in $\mathcal{H}$ that ϵ-shatters the set R and $h_{\overline{R}}$ be the function that shatters the complement of set R.

$$\begin{aligned}
&\geq \frac{1}{2} \mathbb{E}_{\sigma \sim \mathrm{Unif}\{\pm 1\}^{n'}} \left[\frac{1}{n'} \sum_{i=1}^{\mathrm{fat}_\epsilon} \left| \sum_{j=1}^{\lceil n/\mathrm{fat}_\epsilon \rceil} \sigma_{(i-1)\mathrm{fat}_\epsilon + j} \right| 2\epsilon \right] \\
&\geq \frac{\epsilon}{n'} \sum_{i=1}^{\mathrm{fat}_\epsilon} \mathbb{E}_{\sigma \sim \mathrm{Unif}\{\pm 1\}^{n'}} \left[\left| \sum_{j=1}^{\lceil n/\mathrm{fat}_\epsilon \rceil} \sigma_{(i-1)\mathrm{fat}_\epsilon + j} \right| \right] \\
&\geq \frac{\epsilon \ \mathrm{fat}_\epsilon}{n'} \sqrt{\frac{\lceil n/\mathrm{fat}_\epsilon \rceil}{2}} && \text{(Khintchine's inequality)} \\
&= \sqrt{\frac{\epsilon^2 \ \mathrm{fat}_\epsilon}{2 \ n'}}.
\end{aligned}$$

We can now conclude that:

$$\mathrm{fat}_\epsilon \leq \frac{2n' \mathcal{R}_{n'}^2(\mathcal{H})}{\epsilon^2} \leq \frac{4n \mathcal{R}_n^2(\mathcal{H})}{\epsilon^2}$$

where last inequality is because Rademacher complexity decreases with increase in number of samples and $n \leq n' \leq 2n$ (because $\epsilon \geq \mathcal{R}_n(\mathcal{H})$ which implies that $\mathrm{fat}_\epsilon < n$). ☐

B Proof of Lemma 2.2

Recall the key lemma used in proving our main result:

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

$$\mathcal{R}_n(\mathcal{L}_\phi(r)) \leq 21\sqrt{6Hr}\ \log^{\frac{3}{2}}(64\, n)\ \mathcal{R}_n(\mathcal{H})$$

As outlined in Section 2, in order to prove the Lemma 2.2, we take the following steps:

1. We use Lemma A.1 (the refined Dudley Integral bound) to bound the Rademacher complexity of the empirically restricted loss class in terms of the L_2-covering numbers of the class.
2. We use smoothness to get an r-dependent bound on the L_2-covering numbers of the empirically restricted loss class in terms of L_∞-covering numbers of the unrestricted hypothesis class. The key to doing this is the Lemma B.1, which follows from the self bounding property, Lemma 2.1.
3. We bound the L_∞-covering numbers of the unrestricted class in terms of its fat-shattering dimension (Lemma A.2), which in turn can be bounded in terms of its Rademacher complexity (Lemma A.3).

We first present Lemma B.1, which follows from Lemma 2.1 and is the key property we actually use. Equipped with this lemma and the results from Appendix A relating the various complexity measures, we then proceed to the main proof of Lemma 2.2.

Lemma B.1. *For any H-smooth non-negative function $f : \mathbb{R} \mapsto \mathbb{R}$ and any $t, r \in \mathbb{R}$ we have that*

$$(f(t) - f(r))^2 \leq 6H(f(t) + f(r))(t - r)^2 \ .$$

Proof. We start by noting that by the mean value theorem for any $t, r \in \mathbb{R}$ there exists s between t and r such that

$$f(t) - f(r) = f'(s)(t - r) \ . \tag{17}$$

By smoothness, we have that

$$|f'(s) - f'(t)| \leq H\,|t - s| \leq H\,|t - r|\,.$$

Hence we see that

$$|f'(s)| \leq |f'(t)| + H\,|t - r|\ . \tag{18}$$

We now consider two cases:

Case I: If $|t - r| \leq \frac{|f'(t)|}{5H}$ then by (18), $|f'(s)| \leq 6/5\,|f'(t)|$, and combining this with (17) we have:

$$(f(t) - f(r))^2 \leq f'(s)^2(t-r)^2 \leq \frac{36}{25}f'(t)^2(t-r)^2\ .$$

But Lemma 2.1 ensures $f'(t)^2 \leq 4Hf(t)$ yielding:

$$\leq \frac{144}{25}Hf(t)(t-r)^2 < 6Hf(t)(t-r)^2\ . \tag{19}$$

Case II: On the other hand, when $|t - r| > \frac{|f'(t)|}{5H}$, we have from (18) that $|f'(s)| \leq 6H\,|t - r|$. Plugging this into (17) yields:

$$\begin{aligned}(f(t) - f(r))^2 &= |f(t) - f(t)| \cdot |f(t) - f(r)| \leq |f(t) - f(r)|\,(|f'(s)|\,|t - r|)\\ &\leq |f(t) - f(r)|\,(6H\,|t-r| \cdot |t-r|) = 6H\,|f(t) - f(r)|\,(t-r)^2\\ &\leq 6H\max\{f(t), f(r)\}(t-r)^2\ . \end{aligned} \tag{20}$$

Combining the two cases, we have from (19) and (20)) and the non-negativity of $f(\cdot)$, that in either case:

$$(f(t) - f(r))^2 \leq 6H\,(f(t) + f(r))\,(t-r)^2\ . \qquad \square$$

Proof of Lemma 2.2. Following the outline above:

Bounding $\mathcal{R}_n(\mathcal{L}_\phi(r))$ in terms of $\mathcal{N}_2(\mathcal{L}_\phi(r))$ Dudley's integral bound lets us bound the Rademacher complexity of a class in terms of its empirical L_2 covering number. Here we use a more refined version of Dudley's integral bound due to Mendelson [24] and more explicitly stated in [31] and included here for completeness as Lemma A.1:

$$\mathcal{R}_n(\mathcal{L}_\phi(r)) \leq \inf_{\alpha>0}\left\{4\alpha + 10\int_\alpha^{\sqrt{br}}\sqrt{\frac{\mathcal{N}_2\left(\mathcal{L}_\phi(r),\epsilon,n\right)}{n}}d\epsilon\right\}\ . \tag{21}$$

Bounding $\mathcal{N}_2(\mathcal{L}_\phi(r))$ in terms of $\mathcal{N}_\infty(\mathcal{H})$ By Lemma B.1 we see that for a non-negative H-smooth function f, we have that $(f(t) - f(r))^2 \leq 6H(f(t) + f(r))(t-r)^2$. Using this inequality, for any sample $(x_1, y_1), \ldots, (x_n, y_n)$:

$$\begin{aligned}\sqrt{\frac{1}{n}\sum_{i=1}^n(\phi(h(z_i), z_i) - \phi(h_\epsilon(z_i), z_i))^2} &\leq \sqrt{\frac{6H}{n}\sum_{i=1}^n\left(\phi(h(z_i), z_i) + \phi(h_\epsilon(z_i), z_i)\right)(h(z_i) - h_\epsilon(z_i))^2}\\ &\leq \sqrt{\frac{6H}{n}\sum_{i=1}^n\left(\phi(h(z_i), z_i) + \phi(h_\epsilon(z_i), z_i)\right)}\sqrt{\max_{i\in[n]}(h(z_i) - h_\epsilon(z_i))^2}\\ &\leq \sqrt{12Hr}\ \max_{i\in[n]}|h(z_i) - h_\epsilon(z_i)|\ .\end{aligned}$$

That is, an empirical L_∞ cover of $\left\{h \in \mathcal{H} : \hat{L}(h) \leq r\right\}$ at radius $\epsilon/\sqrt{12Hr}$ is also an empirical L_2 cover of $\mathcal{L}_\phi(r)$ at radius ϵ, and we can conclude that:

$$\mathcal{N}_2\left(\mathcal{L}_\phi(r), \epsilon, n\right) \leq \mathcal{N}_\infty\left(\left\{h \in \mathcal{H} : \hat{L}(h) \leq r\right\}, \frac{\epsilon}{\sqrt{12Hr}}, n\right) \leq \mathcal{N}_\infty\left(\mathcal{H}, \frac{\epsilon}{\sqrt{12Hr}}, n\right)\ . \tag{22}$$

Bounding $\mathcal{N}_\infty(\mathcal{H})$ in terms of $\mathcal{R}_n(\mathcal{H})$ Note that for any $\epsilon > \mathcal{R}_n(\mathcal{H})$, by Lemma A.3, $\mathrm{fat}_\epsilon \le n$. Hence the L_∞ covering number at scale $\epsilon/\sqrt{12Hr}$ can be bounded in terms of the fat shattering dimension at that scale using Lemma A.2 as:

$$\mathcal{N}_\infty\left(\mathcal{H}, \frac{\epsilon}{\sqrt{12Hr}}, n\right) \le \left(\frac{2en\,\sqrt{12Hr}B}{\epsilon\ \mathrm{fat}_{\frac{\epsilon}{\sqrt{12Hr}}}(\mathcal{H})}\right)^{\mathrm{fat}_{\frac{\epsilon}{\sqrt{12Hr}}}(\mathcal{H})} . \tag{23}$$

Hence by (21), we have:

$$\mathcal{R}_n(\mathcal{L}_\phi(r)) \le 4\sqrt{12Hr}\mathcal{R}_n(\mathcal{H}) + 10\int_{\sqrt{12Hr}\mathcal{R}_n(\mathcal{H})}^{\sqrt{br}} \sqrt{\frac{\mathrm{fat}_{\frac{\epsilon}{\sqrt{12Hr}}}(\mathcal{H})\log\left(\frac{2en\,\sqrt{12Hr}B}{\epsilon \mathrm{fat}_{\frac{\epsilon}{\sqrt{12Hr}}}(\mathcal{H})}\right)}{n}}d\epsilon$$

and, after a change of integration variable, we have:

$$\begin{aligned}
&\le 4\sqrt{12Hr}\mathcal{R}_n(\mathcal{H}) + 10\sqrt{12Hr}\int_{\mathcal{R}_n(\mathcal{H})}^{\sqrt{b/12H}} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{2en\ B}{\epsilon \mathrm{fat}_\epsilon(\mathcal{H})}\right)}{n}}d\epsilon \\
&\le 4\sqrt{12Hr}\mathcal{R}_n(\mathcal{H}) + 10\sqrt{12Hr}\int_{\mathcal{R}_n(\mathcal{H})}^{\sqrt{b/12H}} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{2eB}{\epsilon}\right)}{n}}d\epsilon \\
&\qquad + 10\sqrt{12Hr}\int_{\mathcal{R}_n(\mathcal{H})}^{\sqrt{b/12H}} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\epsilon(\mathcal{H})}\right)}{n}}d\epsilon .
\end{aligned} \tag{24}$$

We now bound the second term in the sum above. To this end note that for any $\epsilon \ge \mathcal{R}_n(\mathcal{H})$, bounding the fat-shattering dimension in terms of the Rademacher complexity (Lemma A.3) we get:

$$10\sqrt{12Hr}\int_{\mathcal{R}_n(\mathcal{H})}^{\sqrt{b/12H}} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{2eB}{\epsilon}\right)}{n}}d\epsilon \tag{25}$$

$$\begin{aligned}
&\le 10\sqrt{12Hr}\ \mathcal{R}_n(\mathcal{H})\int_{\mathcal{R}_n(\mathcal{H})}^{\sqrt{b/12H}} \frac{\sqrt{\log\left(\frac{2eB}{\epsilon}\right)}}{\epsilon}d\epsilon \\
&\le 10\sqrt{12Hr}\ \mathcal{R}_n(\mathcal{H})\left[-\frac{2}{3}\log^{3/2}\left(\frac{2eB}{\epsilon}\right)\right]_{\mathcal{R}_n(\mathcal{H})}^{\sqrt{b/12H}} \\
&\le \frac{20}{3}\sqrt{12Hr}\ \mathcal{R}_n(\mathcal{H})\left(\log^{3/2}\left(\frac{2eB}{\mathcal{R}_n(\mathcal{H})}\right) - \log^{3/2}\left(\sqrt{\frac{24eHB^2}{b}}\right)\right) \\
&\le \frac{20}{3}\sqrt{12Hr}\ \log^{3/2}\left(\frac{2eB}{\mathcal{R}_n(\mathcal{H})}\right)\mathcal{R}_n(\mathcal{H}) .
\end{aligned} \tag{26}$$

Now we move to the third term of (24), we further split this integral into three parts as:

$$\int_{\mathcal{R}_n(\mathcal{H})}^{\sqrt{b/12H}} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\epsilon(\mathcal{H})}\right)}{n}}d\epsilon \tag{27}$$

$$\leq \int_{\mathcal{R}_n(\mathcal{H})}^{\gamma} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\epsilon(\mathcal{H})}\right)}{n}}d\epsilon + \int_{\gamma}^{\theta} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\epsilon(\mathcal{H})}\right)}{n}}d\epsilon$$
$$+ \int_{\theta}^{\sqrt{b/12H}} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\epsilon(\mathcal{H})}\right)}{n}}d\epsilon\ .$$

Now let θ be such that $\mathrm{fat}_\theta > n/e$, so that for all $\epsilon > \theta$, $\log(n/\mathrm{fat}_\epsilon) \leq 1$. Hence,

$$\leq \int_{\mathcal{R}_n(\mathcal{H})}^{\gamma} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\epsilon(\mathcal{H})}\right)}{n}}d\epsilon + \int_{\gamma}^{\theta} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\epsilon(\mathcal{H})}\right)}{n}}d\epsilon$$
$$+ \int_{\theta}^{\sqrt{b/12H}} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})}{n}}d\epsilon\ .$$

Now to handle the second term in the integral note that in the range $d \in [1, n/e]$, the function $d\log\left(\frac{n}{d}\right)$ is monotonically increasing in d and so in the range of $\epsilon \in [\gamma, \theta]$, $\mathrm{fat}_\epsilon \log\left(\frac{n}{\mathrm{fat}_\epsilon}\right) \leq \mathrm{fat}_\gamma \log\left(\frac{n}{\mathrm{fat}_\gamma}\right)$. Thus we have that

$$\leq \int_{\mathcal{R}_n(\mathcal{H})}^{\gamma} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\epsilon(\mathcal{H})}\right)}{n}}d\epsilon + \int_{\gamma}^{\theta} \sqrt{\frac{\mathrm{fat}_\gamma(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\gamma(\mathcal{H})}\right)}{n}}d\epsilon$$
$$+ \int_{\theta}^{\sqrt{b/12H}} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})}{n}}d\epsilon\ .$$

Further since for all $\epsilon \in [\mathcal{R}_n(\mathcal{H}), \gamma]$ $\mathrm{fat}_\epsilon \leq \mathrm{fat}_\gamma$ we have that

$$\leq \int_{\mathcal{R}_n(\mathcal{H})}^{\gamma} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\gamma(\mathcal{H})}\right)}{n}}d\epsilon + \int_{\gamma}^{\theta} \sqrt{\frac{\mathrm{fat}_\gamma(\mathcal{H})\log\left(\frac{n}{\mathrm{fat}_\gamma(\mathcal{H})}\right)}{n}}d\epsilon$$
$$+ \int_{\theta}^{\sqrt{b/12H}} \sqrt{\frac{\mathrm{fat}_\epsilon(\mathcal{H})}{n}}d\epsilon\ .$$

Since all three integrals above are in the range such that $\epsilon > \mathcal{R}_n(\mathcal{H})$, bounding the fat-shattering dimension in terms of the Rademacher complexity (Lemma A.3) in the first and third integrals :

$$
\begin{aligned}
&\leq \mathcal{R}_n(\mathcal{H}) \int_{\mathcal{R}_n(\mathcal{H})}^{\gamma} \frac{\sqrt{\log\left(\frac{n}{\mathrm{fat}_\gamma(\mathcal{H})}\right)}}{\epsilon} d\epsilon + \int_{\gamma}^{\theta} \sqrt{\frac{\mathrm{fat}_\gamma(\mathcal{H}) \log\left(\frac{n}{\mathrm{fat}_\gamma(\mathcal{H})}\right)}{n}} d\epsilon \\
&\quad + \mathcal{R}_n(\mathcal{H}) \int_{\theta}^{\sqrt{b/12H}} \frac{1}{\epsilon} d\epsilon \\
&\leq \mathcal{R}_n(\mathcal{H}) \sqrt{\log\left(\frac{n}{\mathrm{fat}_\gamma(\mathcal{H})}\right)} \log\left(\frac{1}{\mathcal{R}_n(\mathcal{H})}\right) + \sqrt{\frac{\mathrm{fat}_\gamma(\mathcal{H}) \log\left(\frac{n}{\mathrm{fat}_\gamma(\mathcal{H})}\right)}{n}} (\gamma - \theta) \\
&\quad + \mathcal{R}_n(\mathcal{H}) \log\left(\frac{1}{\theta}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathcal{R}_n(\mathcal{H}) \sqrt{\log\left(\frac{n}{\mathrm{fat}_\gamma(\mathcal{H})}\right)} \log\left(\frac{1}{\mathcal{R}_n(\mathcal{H})}\right) + \sqrt{\frac{\mathrm{fat}_\gamma(\mathcal{H}) \log\left(\frac{n}{\mathrm{fat}_\gamma(\mathcal{H})}\right)}{n}} \sqrt{\frac{b}{12H}} \\
&\quad + \mathcal{R}_n(\mathcal{H}) \log\left(\frac{1}{\mathcal{R}_n(\mathcal{H})}\right)
\end{aligned}
$$

where in the last inequality we used the fact that $\gamma - \theta \leq \sqrt{b/12H}$ (integral range) and that $\theta \geq \mathcal{R}_n(\mathcal{H})$. Picking γ to be such that $\mathrm{fat}_\gamma = 12Hn\mathcal{R}_n^2(\mathcal{H})/b$ we conclude that

$$
\begin{aligned}
&\leq \mathcal{R}_n(\mathcal{H}) \sqrt{\log\left(\frac{b}{12H\mathcal{R}_n^2(\mathcal{H})}\right)} \log\left(\frac{1}{\mathcal{R}_n(\mathcal{H})}\right) + \mathcal{R}_n(\mathcal{H}) \sqrt{\log\left(\frac{b}{12H\mathcal{R}_n^2(\mathcal{H})}\right)} \\
&\quad + \mathcal{R}_n(\mathcal{H}) \log\left(\frac{1}{\mathcal{R}_n(\mathcal{H})}\right) \\
&\leq 3\ \mathcal{R}_n(\mathcal{H}) \log^{3/2}\left(\frac{2eB}{\mathcal{R}_n(\mathcal{H})}\right) \ . \qquad (28)
\end{aligned}
$$

Hence plugging back the above and (25) back in (24) we conclude that

$$
\begin{aligned}
\mathcal{R}_n(\mathcal{L}_\phi(r)) &\leq 4\sqrt{12Hr}\mathcal{R}_n(\mathcal{H}) + 7\sqrt{12Hr}\mathcal{R}_n(\mathcal{H}) \log^{3/2}\left(\frac{2eB}{\mathcal{R}_n(\mathcal{H})}\right) \qquad (29) \\
&\quad + 30\sqrt{12Hr}\mathcal{R}_n(\mathcal{H}) \log^{3/2}\left(\frac{2eB}{\mathcal{R}_n(\mathcal{H})}\right) \\
&\leq 41\sqrt{12Hr}\mathcal{R}_n(\mathcal{H}) \log^{3/2}\left(\frac{2eB}{\mathcal{R}_n(\mathcal{H})}\right) \ . \qquad (30)
\end{aligned}
$$

Now by definition of Rademacher complexity, we have,

$$\begin{aligned}
\mathcal{R}_n(\mathcal{H}) &= \sup_{x_1,\ldots,x_n \in \mathcal{X}} \mathbb{E}_{\sigma \sim \mathrm{Unif}(\{\pm 1\}^n)} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \left| \sum_{i=1}^{n} h(x_i)\sigma_i \right| \right] \\
&\geq \sup_{x \in \mathcal{X}} \mathbb{E}_{\sigma \sim \mathrm{Unif}(\{\pm 1\}^n)} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \left| \sum_{i=1}^{n} h(x)\sigma_i \right| \right] \\
&= \left(\sup_{x \in \mathcal{X}} \sup_{h \in \mathcal{H}} |h(x)| \right) \left(\mathbb{E}_{\sigma \sim \mathrm{Unif}(\{\pm 1\}^n)} \left[\frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \right| \right] \right) \\
&= B \ \mathbb{E}_{\sigma \sim \mathrm{Unif}(\{\pm 1\}^n)} \left[\frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \right| \right] \geq \frac{B}{\sqrt{2n}}
\end{aligned}$$

where the last step is due to Khintchine's inequality (see, e.g., page 364 of [9]). Thus we see that $\frac{2eB}{\mathcal{R}_n(\mathcal{H}} \leq 8\sqrt{n}$. Plugging this in (29), we conclude that

$$\mathcal{R}_n(\mathcal{L}_\phi(r)) \leq 21\sqrt{6Hr}\ \log^{\frac{3}{2}}(64\ n)\ \mathcal{R}_n(\mathcal{H})\ . \qquad \square$$