Journal of Machine Learning Research 16 (2015) 3475-3486 Submitted 9/14; Revised 3/15; Published 12/15

The Sample Complexity of Learning Linear Predictors with the Squared Loss

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
Rehovot 7610001, Israel

Editor: Mehryar Mohri

Abstract

We provide a tight sample complexity bound for learning bounded-norm linear predictors with respect to the squared loss. Our focus is on an agnostic PAC-style setting, where no assumptions are made on the data distribution beyond boundedness. This contrasts with existing results in the literature, which rely on other distributional assumptions, refer to specific parameter settings, or use other performance measures.

Keywords: sample complexity, squared loss, linear predictors, distribution-free learning

1. Introduction

In machine learning and statistics, the squared loss is the most commonly used loss for measuring real-valued predictions: Given a prediction p and actual target value y, it is defined as $\ell(p, y) = (p - y)^2$. It is intuitive, has a convenient analytical form, and has been extremely well-studied.

In this paper, we concern ourselves with learning bounded-norm linear predictors with respect to the squared loss, in an agnostic PAC learning framework. Formally, for some fixed parameters X, Y, B, we assume the existence of an unknown distribution over $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq X\} \times \{y \in \mathbb{R} : |y| \leq Y\}$, from which we are given a training set $S = \{\mathbf{x}_i, y_i\}_{i=1}^m$ of m i.i.d. examples, consisting of pairs of instances $\mathbf{x}$ and target values y. Given a linear predictor $\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$, its risk with respect to the squared loss is defined as

$$R(\mathbf{w}) \;=\; \mathbb{E}\left[(\langle \mathbf{w}, \mathbf{x} \rangle - y)^2\right],$$

where the expectation is with respect to $\mathbf{x}, y$. Our goal is to find a linear predictor $\mathbf{w}$ from the hypothesis class of norm-bounded linear predictors,

$$\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\| \leq B\},$$

such that its excess risk

$$R(\mathbf{w}) - \min_{\mathbf{w} \in \mathcal{W}} R(\mathbf{w})$$

with respect to the best possible predictor in $\mathcal{W}$ is as small as possible. We focus here on the expected excess risk (over the randomness of the training set and algorithm), and study what is the optimal bound on the excess risk one can obtain—also known as a sample complexity

bound—and how it is affected by the problem parameters X, Y, B, d and the sample size m, uniformly over any distribution. Since X and B are invariant to simultaneous scaling (if we re-scale each $\mathbf{x}$ by some factor c, and re-scale each linear predictor $\mathbf{w}$ by $1/c$, all predictions remain the same), we will assume without loss of generality that $X = 1$.

There is a huge literature on learning with the squared loss, with many tight and elegant risk bounds under various assumptions. However, for the framework defined above, there does not appear to be an explicit and self-contained analysis. Much of the existing work (some examples include Hsu et al. (2014); Koltchinskii (2011); Lecué and Mendelson (2014); Tsybakov (2003); Anthony and Bartlett (1999); Lee et al. (1998); Zhang (2005); Audibert and Catoni (2011)) focuses on risk upper bounds, but not lower bounds showing the limits of attainable performance. Moreover, most existing work considers settings different than ours in one or more of the following aspects:

- *Additional Distributional Assumptions:* In our agnostic setting, we make no assumptions on the data distribution except boundedness. In contrast, most existing work relies on additional assumptions. Perhaps the most common assumption is a well-specified model, under which there exists a fixed $\mathbf{w} \in \mathbb{R}^d$ such that $y = \langle \mathbf{w}, \mathbf{x} \rangle + \xi$, where ξ is a zero-mean noise term (such as Tsybakov (2003)). Other works impose additional conditions on the distribution of $\mathbf{x}$ (for example, that the covariance matrix of $\mathbf{x}$ is well behaved, such as Hsu et al. (2014)), or consider a fixed design setting where the data instances $\mathbf{x}$ are not sampled i.i.d.. These assumptions usually lead to excess risk bounds which scale (at least in finite dimensions) as dY^2/m, independent of the norm bound B. However, as we will see later, this is not the behavior in our setting.

- *Bounds not on the excess risk:* Many of the existing results are not on the excess risk, but rather on $\mathbb{E}[\|\mathbf{w} - \mathbf{w}^*\|^2]$ or $\mathbb{E}[(\langle \mathbf{w}, \mathbf{x} \rangle - \langle \mathbf{w}^*, \mathbf{x} \rangle)^2]$, where $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} R(\mathbf{w})$ (such as Koltchinskii (2011); Lecué and Mendelson (2014)). The former measure is relevant for parameter estimation, while the latter measure can be shown to equal the excess risk when $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^d} R(\mathbf{w})$ (in other words, $B = \infty$, see Lemma 2 below). However, when we deal with the hypothesis class of norm-bounded predictors, then the excess risk can be larger by an arbitrary factor[1]. Therefore, upper bounds on these measures do not imply upper bounds on the excess risk in our setting. We remark that in our distribution-free setting, we must constrain the hypothesis class, since if our hypothesis class contains all linear predictors ($B = \infty$), then the lower bounds below imply that non-trivial learning is impossible with any sample size (regardless of the dimension d).

- *Bounded Functions:* Many learning theory results for the squared loss (such as those based on fat-shattering techniques, see Lee et al. (1998); Anthony and Bartlett (1999)) assume that the predictor functions and target values are bounded in some fixed

1. For example, consider a distribution on (x, y) such that $(x, y) = (1, 1)$ with probability 1, and $\mathcal{W} = \{w : w \in [-1/2, 1/2]\}$. Then clearly, $w^* = 1/2$, and $\mathbb{E}[(wx - w^*x)^2] = \mathbb{E}[(w - w^*)^2] = (1/2 - w)^2$. However, the excess risk equals $(w-1)^2 - (1/2-1)^2 = w^2 - 2w + 3/4 = (1/2-w)^2 + (1/2-w)$. This is larger than the excess risk by an additive factor of $(1/2 - w)$, and a multiplicative factor of $\frac{1}{1/2-w}$—arbitrarily large if w is close to $w^* = 1/2$.

interval (such as [-1,+1]). In our setting, this would correspond to assuming $B, Y \leq 1$. Other results (such as Bartlett and Mendelson (2003)) assume Lipschitz loss functions, which is not satisfied for the squared loss. One notable exception is Srebro et al. (2010), which analyzes smooth and strongly-convex losses (such as the squared loss) and provide tight sample complexity bounds. However, their results apply either when the functions are bounded by 1, or when d is extremely large or infinite dimensional. In contrast, we provide more general results which hold for any d and when the functions are not necessarily bounded by 1.

- *Collapsing Problem Parameters Together:* Some works, such as Srebro et al. (2010), implicitly take Y to equal the largest possible prediction, $\sup_{\mathbf{w},\mathbf{x}} |\langle \mathbf{w}, \mathbf{x} \rangle| = B$, and give results only in terms of B. However, we will see that B and Y affect the excess risk in a different manner, and it is thus important to discern between them. Moreover, B and Y can often have very different magnitudes. For example, in learning problems where the instances $\mathbf{x}$ tend to be sparse, we may want to have the norm bound B of the predictor to scale with the dimension d, while the bound on the target values Y remain a fixed constant.

2. Main Result

Our main result is the following lower bound on the attainable excess risk:

Theorem 1 *There exists a universal constant c, such that for any dimension d, sample size m, target value bound Y, predictor norm bound $B \geq 2Y$, and for any algorithm returning a linear predictor $\hat{\mathbf{w}}$, there exists a data distribution such that*

$$\mathbb{E}[R(\hat{\mathbf{w}}) - R(\mathbf{w}^*)] \geq c \ \min\left\{ Y^2, \frac{B^2 + dY^2}{m}, \frac{BY}{\sqrt{m}} \right\},$$

where $\mathbf{w}^* = \arg\min_{\mathbf{w}:\|\mathbf{w}\| \leq B} R(\mathbf{w})$, and the expectation is with respect to the training set and the (possible) randomness of the algorithm.

Based on existing results in the literature, this bound has essentially matching upper bounds, up to logarithmic factors:

- Using the trivial zero predictor $\hat{\mathbf{w}} = \mathbf{0}$, we are guaranteed that $R(\hat{\mathbf{w}}) - R(\mathbf{w}^*) \leq R(\hat{\mathbf{w}}) = \mathbb{E}[(\langle \mathbf{0}, \mathbf{x} \rangle - y)^2] = \mathbb{E}[y^2] \leq Y^2$.
- Using the Vovk-Azoury-Warmuth forecaster (Vovk (2001); Azoury and Warmuth (2001)) and a standard online-to-batch conversion technique (see for instance Shalev-Shwartz (2012), corollary 5.2), we have an algorithm for which

$$\mathbb{E}[R(\hat{\mathbf{w}}) - R(\mathbf{w}^*)] \leq \mathcal{O}\left(\frac{B^2 + dY^2 \log(1 + m/d)}{m} \right).$$

- Alternatively, by corollary 3 in Srebro et al. (2010) [2], using mirror descent with an online-to-batch conversion gives us an algorithm for which $\mathbb{E}[R(\hat{\mathbf{w}}) - R(\mathbf{w}^*)] \leq$

2. Where $\bar{L}^* \leq Y^2$ and $H = 2$ for the squared loss.

$\mathcal{O}\left(\frac{BY}{\sqrt{m}} + \frac{B^2}{m}\right)$. In the regime where this bound is smaller than Y^2, it can be verified that $BY/\sqrt{m}$ is the dominant term, in which case we get an $\mathcal{O}(BY/\sqrt{m})$ bound.

Taking the best of these algorithmic approaches, we get the minimum of these upper bounds, i.e. we can find a predictor $\hat{\mathbf{w}}$ for which

$$\mathbb{E}[R(\hat{\mathbf{w}}) - R(\mathbf{w}^*)] \leq \mathcal{O}\left(\min\left\{Y^2, \frac{B^2 + dY^2 \log\left(1 + \frac{m}{d}\right)}{m}, \frac{BY}{\sqrt{m}}\right\}\right).$$

We conjecture that the same bound, perhaps up to log-factors, can be shown for empirical risk minimization (i.e. given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, return $\hat{\mathbf{w}} = \min_{\mathbf{w}:\|\mathbf{w}\|\leq B} \frac{1}{m}\sum_{i=1}^m (\langle \mathbf{w}, \mathbf{x}_i\rangle - y_i)^2$).

This result has some interesting consequences: First, it implies that even when $d = 1$ (i.e. a one-dimensional problem), there is a non-trivial dependence on the norm bound B. This is in contrast to results under the well-specified model or other common distributional assumptions, which lead to upper bounds independent of B. Second, it shows that in a finite-dimensional setting, although the squared loss $(\langle \mathbf{w}, \mathbf{x}\rangle - y)^2$ may appear symmetric with respect to y and $\langle \mathbf{w}, \mathbf{x}\rangle$, the attainable excess risk is actually much more sensitive to the bound Y on $|y|$ than to the bound B on $|\langle \mathbf{w}, \mathbf{x}\rangle|$, due to the d factor. For example, if Y is a constant, then B can be as large as the dimension d without affecting the leading term of the excess risk. Third, in the context of online learning, it implies that the Vovk-Azoury-Warmuth forecaster is essentially optimal in our setting and for a finite-dimensional regime, in terms of its dependence on both d and B (the lower bounds in Vovk (2001); Singer et al. (2002) do not show an explicit dependence on B).

3. Proof of Thm. 1

The proof of our main result consist of two separate lower bounds, each of which uses a different construction. The theorem follows by combining them and performing a few simplifications.

We begin by recalling the following result, which follows from the well-known orthogonality principle:

Lemma 2 *Let $R(\mathbf{w}) = \mathbb{E}[(\langle \mathbf{w}, \mathbf{x}\rangle - y)^2]$, where the expectation is over $\mathbf{x}, y$, and let $\mathbf{w}^* = \arg\min_{\mathbf{w}:\|\mathbf{w}\|\leq B} R(\mathbf{w})$. Then for any $\mathbf{w} : \|\mathbf{w}\| \leq B$, it holds that*

$$R(\mathbf{w}) - R(\mathbf{w}^*) \geq \mathbb{E}[(\langle \mathbf{w}, \mathbf{x}\rangle - \langle \mathbf{w}^*, \mathbf{x}\rangle)^2],$$

with equality when $B = \infty$.

Proof For any $\mathbf{w} \in \mathbb{R}^d$, define the linear function $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ by $f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle$. Then $\{f_{\mathbf{w}}(\cdot) : \|\mathbf{w}\| \leq B\}$ corresponds to a closed convex set in the L^2 function space defined via the inner product $\langle f, g\rangle = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})g(\mathbf{x})]$ and norm $\|f\|^2 = \mathbb{E}_{\mathbf{x}}[f^2(\mathbf{x})]$. Moreover, letting $\eta(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$, we have

$$\begin{aligned} R(\mathbf{w}) - R(\mathbf{w}^*) &= \mathbb{E}[(\langle \mathbf{w}, \mathbf{x}\rangle - y)^2] - \mathbb{E}[(\langle \mathbf{w}, \mathbf{x}\rangle - y)^2] \\ &= \mathbb{E}[(f_{\mathbf{w}}(\mathbf{x}) - \eta(\mathbf{x}))^2] - \mathbb{E}[(f_{\mathbf{w}^*}(\mathbf{x}) - \eta(\mathbf{x}))^2] \\ &= \|f_{\mathbf{w}} - \eta\|^2 - \|f_{\mathbf{w}^*} - \eta\|^2. \end{aligned}$$

Moreover,

$$\mathbb{E}[(\langle \mathbf{w}, \mathbf{x}\rangle - \langle \mathbf{w}^*, \mathbf{x}\rangle)^2] = \mathbb{E}[\langle \mathbf{w} - \mathbf{w}^*, \mathbf{x}\rangle^2] = \|f_{\mathbf{w}-\mathbf{w}^*}\|^2 = \|f_{\mathbf{w}} - f_{\mathbf{w}^*}\|^2.$$

Therefore, the inequality in the lemma can be written as

$$\|f_{\mathbf{w}} - \eta\|^2 - \|f_{\mathbf{w}^*} - \eta\|^2 \;\geq\; \|f_{\mathbf{w}} - f_{\mathbf{w}^*}\|^2,$$

or equivalently

$$\|f_{\mathbf{w}} - f_{\mathbf{w}^*}\|^2 + \|f_{\mathbf{w}^*} - \eta\|^2 \;\leq\; \|f_{\mathbf{w}} - \eta\|^2. \tag{1}$$

To see why this inequality hold, recall that the set of linear functionals (which includes $f_{\mathbf{w}}$ and $f_{\mathbf{w}^*}$) form a linear subspace in L^2. Moreover, $f_{\mathbf{w}^*}$ is the projection of η on the set $\{f_{\mathbf{w}} : \|\mathbf{w}\| \leq B\}$: To see this, note that

$$\begin{aligned}
\mathbf{w}^* &= \arg\min_{\mathbf{w}:\|\mathbf{w}\|\leq B} \mathbb{E}[(\langle \mathbf{w}, \mathbf{x}\rangle - y)^2 \\
&= \arg\min_{\mathbf{w}:\|\mathbf{w}\|\leq B} \mathbb{E}\left[\mathbb{E}\left[(\langle \mathbf{w}, \mathbf{x}\rangle - y)^2 | \mathbf{x}\right]\right] \\
&= \arg\min_{\mathbf{w}:\|\mathbf{w}\|\leq B} \mathbb{E}\left[\langle \mathbf{w}, \mathbf{x}\rangle^2 - 2\mathbb{E}\left[\langle \mathbf{w}, y\mathbf{x}\rangle | \mathbf{x}\right] + \mathbb{E}[y^2|\mathbf{x}] \mid \mathbf{x}\right] \\
&= \arg\min_{\mathbf{w}:\|\mathbf{w}\|\leq B} \mathbb{E}\left[\langle \mathbf{w}, \mathbf{x}\rangle^2 - 2\langle \mathbf{w}, \mathbb{E}[y|\mathbf{x}]\mathbf{x}\rangle + \mathbb{E}_y[y|\mathbf{x}]^2 \mid \mathbf{x}\right] \\
&= \arg\min_{\mathbf{w}:\|\mathbf{w}\|\leq B} \mathbb{E}\left[\left(\langle \mathbf{w}, \mathbf{x}\rangle - \eta(\mathbf{x})\right)^2\right],
\end{aligned}$$

(where in the fourth equality we used the fact that adding and subtracting terms independent of $\mathbf{w}$ does not change the argmin), and therefore $f_{\mathbf{w}^*} = \arg\min_{f_{\mathbf{w}}:\|\mathbf{w}\|\leq B} \|f_{\mathbf{w}} - \eta\|^2$. When $B = \infty$, then $f_{\mathbf{w}^*}$ is simply the projection of η on the linear sub-space of linear functionals, hence (1) holds with equality by the Pythagorean theorem (see figure 1). When $B < \infty$, then $f_{\mathbf{w}^*}$ is the projection of η on a constrained convex subset of this linear space, and we only have an inequality. ■

Our first construction provides an excess risk lower bound even when we deal with one-dimensional problems:

Theorem 3 *There exists a universal constant c, such that for any sample size m, target value bound Y, predictor norm bound $B \geq 2Y$, and any algorithm returning a linear predictor $\hat{\mathbf{w}}$, there exists a data distribution in $d = 1$ dimensions such that*

$$\mathbb{E}[R(\hat{w}) - R(w^*)] \geq c \min\left\{Y^2, \frac{B^2}{m}\right\}.$$

The expectation is with respect to the training set and the (possible) randomness of the algorithm.

Proof Let α, γ be small positive parameters in $(0, 1]$ to be chosen later, such that $\alpha > \gamma$, and consider the following two distributions over (x, y):

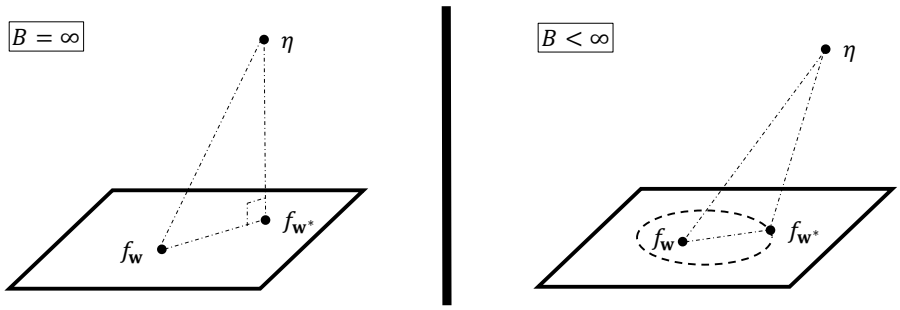

Figure 1: Illustration of inequality in the proof of Lemma 2. The rectangle represents the subspace of linear functionals, and the dotted circle in the right figure represents the convex subset $\{f_{\mathbf{w}} : \|\mathbf{w}\| \leq B\}$.

- Distribution $\mathcal{D}_0$: $y = Y$ w.p. 1; $x = \begin{cases} Y/B & \text{w.p. } \alpha \\ 0 & \text{w.p. } 1-\alpha \end{cases}$.
- Distribution $\mathcal{D}_1$: $y = Y$ w.p. 1; $x = \begin{cases} 1 & \text{w.p. } \gamma \\ Y/B & \text{w.p. } \alpha - \gamma \\ 0 & \text{w.p. } 1-\alpha \end{cases}$.

Note that since $B \geq 2Y$, $|x| \leq 1$, so these are indeed valid distributions. Intuitively, in both distributions x is small most of the time, but under $\mathcal{D}_1$ it can occasionally have a "large" value of 1. Unless the sample size is large enough, it is not possible to distinguish between these two distributions, and this will lead to an excess risk lower bound.

Let $\mathbb{E}_0$ and $\mathbb{E}_1$ denote expectations with respect to $\mathcal{D}_0$ and $\mathcal{D}_1$ respectively. Let

$$w_0^* = B$$

denote the optimal predictor under $\mathcal{D}_0$, and let

$$w_1^* = \frac{\mathbb{E}_1[yx]}{\mathbb{E}_1[x^2]} = \frac{(Y^2/B)(\alpha-\gamma)+Y\gamma}{(Y^2/B^2)(\alpha-\gamma)+\gamma} = B\frac{Y^2(\alpha-\gamma)+BY\gamma}{Y^2(\alpha-\gamma)+B^2\gamma}$$

denote the optimal predictor under $\mathcal{D}_1$. Note that since $B \geq 2Y$, we have $w_1^* \leq w_0^*$, and moreover,

$$\begin{aligned}(w_1^* - w_0^*)^2 = B^2\left(\frac{Y^2(\alpha-\gamma)+BY\gamma}{Y^2(\alpha-\gamma)+B^2\gamma} - 1\right)^2 &= B^4\gamma^2\left(\frac{Y-B}{Y^2\alpha+(B^2-Y^2)\gamma}\right)^2 \\ &\geq B^4\gamma^2\left(\frac{Y-B}{Y^2\alpha+B^2\gamma}\right)^2. \qquad (2)\end{aligned}$$

By Yao's minimax principle, it is sufficient to show that when choosing either $\mathcal{D}_0$ or $\mathcal{D}_1$ uniformly at random, and generating a dataset according to that distribution, any

deterministic algorithm attains the lower bound in the theorem. Using Lemma 2, and the notation $\Pr_0$ (respectively $\Pr_1$) to denote probabilities with respect to $\mathcal{D}_0$ (respectively $\mathcal{D}_1$), we have

$$\begin{aligned}
\mathbb{E}[R(\hat{w})-R(w^*)] &= \frac{1}{2}\left(\mathbb{E}_0[(\hat{w}x - w_0^*x)^2] + \mathbb{E}_1[(\hat{w}x - w_1^*x)^2]\right) \\
&\geq \frac{1}{2}\frac{Y^2\alpha}{B^2}\left(\mathbb{E}_0[(\hat{w} - w_0^*)^2] + \mathbb{E}_1[(\hat{w} - w_1^*)^2]\right) \\
&\geq \frac{1}{2}\frac{Y^2\alpha}{B^2}\left(\frac{w_1^* - w_0^*}{2}\right)^2\left(\Pr_0\left(\hat{w} < \frac{w_0^*+w_1^*}{2}\right) + \Pr_1\left(\hat{w} \geq \frac{w_0^*+w_1^*}{2}\right)\right) \\
&= \frac{1}{2}\frac{Y^2\alpha}{B^2}\left(\frac{w_1^* - w_0^*}{2}\right)^2\left(1 - \left(\Pr_0\left(\hat{w} \geq \frac{w_0^*+w_1^*}{2}\right) - \Pr_1\left(\hat{w} \geq \frac{w_0^*+w_1^*}{2}\right)\right)\right) \\
&\geq \frac{1}{2}\frac{Y^2\alpha}{B^2}\left(\frac{w_1^* - w_0^*}{2}\right)^2\left(1 - \left|\Pr_0\left(\hat{w} \geq \frac{w_0^*+w_1^*}{2}\right) - \Pr_1\left(\hat{w} \geq \frac{w_0^*+w_1^*}{2}\right)\right|\right),
\end{aligned}$$

where in the second inequality we used the fact that $w_1^* \leq w_0^*$. By Pinsker's inequality, since $\hat{w}$ is a deterministic function of the training set S, this is at least

$$\frac{1}{8}\frac{Y^2\alpha}{B^2}\left(w_1^* - w_0^*\right)^2\left(1 - \sqrt{\frac{1}{2}D_{kl}(p_0(S)||p_1(S))}\right),$$

where D_{kl} is the Kullback-Leibler divergence, and p_0 (respectively p_1) is the probability measure of the sample with respect to $\mathcal{D}_0$ (respectively $\mathcal{D}_1$). Since S is composed of m i.i.d. instances, and the target value y is fixed under both distributions, we can invoke the chain rule and rewrite this as

$$\frac{1}{8}\frac{Y^2\alpha}{B^2}\left(w_1^* - w_0^*\right)^2\left(1 - \sqrt{\frac{m}{2}\ D_{kl}(p_0(x)||p_1(x))}\right).$$

To simplify the bound, we use the following fact (see for instance Gibbs and Su (2002), Theorem 5):

Lemma 4 *For any probability distributions p, q over the same discrete sample space, it holds that $D_{kl}(p||q)$ is upper bounded by the χ^2 divergence between p and q, which equals $\sum_a \frac{(p(a)-q(a))^2}{q(a)}$.*

Using this lemma, we have

$$D_{kl}(p_0(x)||p_1(x)) \leq \frac{\gamma^2}{\gamma} + \frac{\gamma^2}{\alpha-\gamma} = \gamma\left(1 + \frac{\gamma}{\alpha-\gamma}\right).$$

Plugging this back, as well as the value of $(w_1^* - w_0^*)^2$ from (2), we get an excess loss lower bound on the form

$$\frac{1}{8}Y^2\alpha B^2\gamma^2\left(\frac{Y-B}{Y^2\alpha + B^2\gamma}\right)^2\left(1 - \sqrt{\frac{m}{2}\gamma\left(1 + \frac{\gamma}{\alpha-\gamma}\right)}\right),$$

We now consider two cases:

- If $m \leq B^2/Y^2$, we pick $\alpha = 1$ and $\gamma = 1/3m$, and get that the expression above is at least

$$\begin{aligned}
&\frac{Y^2}{72}\frac{B^2}{m^2}\left(\frac{B-Y}{Y^2+B^2/3m}\right)^2\left(1-\sqrt{\frac{1}{6}\left(1+\frac{1/3m}{1-1/3m}\right)}\right) \\
&= \frac{Y^2}{72}\left(\frac{B(B-Y)}{mY^2+B^2/3}\right)^2\left(1-\sqrt{\frac{1}{6}\left(1+\frac{1}{3m-1}\right)}\right) \\
&\geq \frac{Y^2}{72}\left(\frac{B(B-Y)}{(B^2/Y^2)Y^2+B^2/3}\right)^2\left(1-\sqrt{\frac{1}{6}\left(1+\frac{1}{3m-1}\right)}\right) \\
&\geq \frac{Y^2}{72}\left(\frac{B(B-Y)}{(1+1/3)B^2}\right)^2\left(1-\sqrt{\frac{1}{6}\left(1+\frac{1}{2}\right)}\right) \\
&\geq 0.003\ Y^2\left(\frac{B-Y}{B}\right)^2 \ = \ 0.003\ Y^2\left(1-\frac{Y}{B}\right)^2 \ \geq \ 0.003\ Y^2\left(1-\frac{1}{2}\right)^2,
\end{aligned}$$

  where we used the assumption that $B \geq 2Y$.

- If $m > B^2/Y^2$, we pick $\alpha = B^2/(Y^2 m)$ and $\gamma = 1/3m$ and get that the expression above is at least

$$\begin{aligned}
&\frac{1}{8}\frac{B^4}{m}\frac{1}{9m^2}\left(\frac{B-Y}{B^2/m+B^2/3m}\right)^2\left(1-\sqrt{\frac{1}{6}\left(1+\frac{1/3m}{(B^2/Y^2-1/3)/m}\right)}\right) \\
&\geq \frac{1}{72}\frac{(B-Y)^2}{m(1+1/3)^2}\left(1-\sqrt{\frac{1}{6}\left(1+\frac{1/3}{4-1/3}\right)}\right) \\
&\geq 0.004\frac{(B-Y)^2}{m} \ \geq \ 0.004\frac{(B-B/2)^2}{m} \ = \ 0.001\frac{B^2}{m},
\end{aligned}$$

  where we used the assumption that $B \geq 2Y$.

Combining the two cases, we get an excess risk lower bound of $c\ \min\left\{Y^2, \frac{B^2}{m}\right\}$ for some universal constant c. ■

Our second construction provides a different type of bound, which quantifies a dependence on the dimension d:

Theorem 5 *There exists a universal constant c, such that for any dimension d, sample size m, target value bound Y, predictor norm bound B and any algorithm returning a linear predictor $\hat{\mathbf{w}}$, there exists a data distribution in d dimensions such that*

$$\mathbb{E}[R(\hat{\mathbf{w}}) - R(\mathbf{w}^*)] \geq c\ \min\left\{Y^2, B^2, \frac{dY^2}{m}, \frac{BY}{\sqrt{m}}\right\}.$$

The expectation is with respect to the training set and the (possible) randomness of the algorithm.

Proof By Yao's minimax principle, it is sufficient to display a randomized choice of data distributions, with respect to which the expected excess error of any deterministic algorithm attains the lower bound in the theorem. In the sequel, we use $\mathbb{E}$ to denote expectation with respect to the random choice of data distribution, as well as the random drawing of a training set from the distribution.

In particular, fix some $d' \leq d$ to be chosen later, let $\boldsymbol{\sigma} \in \{-1,+1\}^{d'}$ be chosen uniformly at random, and consider the distribution $\mathcal{D}_{\boldsymbol{\sigma}}$ (indexed by $\boldsymbol{\sigma}$) over examples $(\mathbf{x}, y)$, defined as follows: $\mathbf{x}$ is chosen uniformly at random among the first d' standard basis vectors. Conditioned on $\mathbf{x} = \mathbf{e}_i$, y is chosen to equal Y with probability $\frac{1}{2}(1+\sigma_i b)$, where $b = \min\{1/2, \sqrt{d'/6m}\}$, and $-Y$ otherwise.

A simple calculation shows that the optimum $\mathbf{w}^* = \arg\min_{\mathbf{w}:\|\mathbf{w}\|\leq B} R(\mathbf{w})$ is such that

$$\forall\, i \in \{1,\ldots,d'\}, \quad w_i^* = \sigma_i \min\{Yb, B/\sqrt{d'}\}.$$

Therefore, using Lemma 2 and the notation $\mathbf{1}_A$ as the indicator function for the event A:

$$\begin{aligned}
\mathbb{E}\left[R(\hat{\mathbf{w}}) - R(\mathbf{w}^*)\right] &\geq \mathbb{E}\left[\mathbb{E}_{\mathbf{x}}[(\langle\hat{\mathbf{w}},\mathbf{x}\rangle - \langle\mathbf{w}^*,\mathbf{x}\rangle)^2]\right] \\
&= \frac{1}{d'}\sum_{i=1}^{d'} \mathbb{E}[(\hat{\mathbf{w}}_i - \mathbf{w}_i^*)^2] \\
&\geq \frac{1}{d'}\sum_{i=1}^{d'} \mathbb{E}[(\mathbf{w}_i^*)^2 \mathbf{1}_{\hat{\mathbf{w}}_i \mathbf{w}_i^* \leq 0}] \\
&= \frac{1}{d'}\left(\min\{Yb, B/\sqrt{d'}\}\right)^2 \sum_{i=1}^{d'} \Pr(\hat{\mathbf{w}}_i \mathbf{w}_i^* \leq 0).
\end{aligned}$$

Since σ_i is uniformly distributed on $\{-1,+1\}$, and has the same sign as w_i^*, this equals

$$\begin{aligned}
&\frac{1}{d'}\left(\min\{Yb, B/\sqrt{d'}\}\right)^2 \sum_{i=1}^{d'} \frac{1}{2}\left(\Pr(\hat{\mathbf{w}}_i \geq 0 | \sigma_i < 0) + \Pr(\hat{\mathbf{w}}_i \leq 0 | \sigma_i > 0)\right) \\
&\geq \frac{1}{2d'}\left(\min\{Yb, B/\sqrt{d'}\}\right)^2 \sum_{i=1}^{d'} \left(1 - \Pr(\hat{\mathbf{w}}_i \leq 0 | \sigma_i < 0) + \Pr(\hat{\mathbf{w}}_i \leq 0 | \sigma_i > 0)\right) \\
&\geq \frac{1}{2d'}\left(\min\{Yb, B/\sqrt{d'}\}\right)^2 \sum_{i=1}^{d'} \left(1 - \left|\Pr(\hat{\mathbf{w}}_i \leq 0 | \sigma_i < 0) - \Pr(\hat{\mathbf{w}}_i \leq 0 | \sigma_i > 0)\right|\right)
\end{aligned}$$

Using Pinsker's inequality and the fact that $\hat{\mathbf{w}}$ is a deterministic function of the training set S, this is at least

$$\frac{1}{2d'}\left(\min\{Yb, B/\sqrt{d'}\}\right)^2 \sum_{i=1}^{d'} \left(1 - \sqrt{\frac{1}{2} D_{kl}\left(p(S|\sigma_i > 0) || p(S|\sigma_i < 0)\right)}\right), \tag{3}$$

where D_{kl} is the Kullback-Leibler (KL) divergence, and p is the probability measure of the sample. Since the training set is composed of m i.i.d. instances, we can use the chain rule

and get that this divergence equals $mD_{kl}\left(p((\mathbf{x},y)|\sigma_i>0)||p((\mathbf{x},y)|\sigma_i<0)\right)$. Moreover, we note that

$$\begin{aligned} p((\mathbf{x},y)|\sigma_i) = &\ p(\mathbf{x}=\mathbf{e}_i)p((\mathbf{x},y)|\sigma_i,\mathbf{x}_i=\mathbf{e}_i)+p(\mathbf{x}\neq\mathbf{e}_i)p((\mathbf{x},y)|\sigma_i,\mathbf{x}\neq\mathbf{e}_i) \\ = &\ \frac{1}{d'}p((\mathbf{x},y)|\sigma_i,\mathbf{x}=\mathbf{e}_i)+\left(1-\frac{1}{d'}\right)p((\mathbf{x},y)|\sigma_i,\mathbf{x}\neq\mathbf{e}_i), \end{aligned}$$

and therefore, by joint convexity of the KL-divergence[3]

$$\begin{aligned} &D_{kl}(p((\mathbf{x},y)|\sigma_i>0)||p(\mathbf{x},y)|\sigma_i<0)) \\ &\qquad \leq \ \frac{1}{d'}D_{kl}\left(p((\mathbf{x},y)|\sigma_i>0,\mathbf{x}=\mathbf{e}_i)||p((\mathbf{x},y)|\sigma_i<0,\mathbf{x}=\mathbf{e}_i)\right) \\ &\qquad\quad +\left(1-\frac{1}{d'}\right)D_{kl}\left(p((\mathbf{x},y)|\sigma_i>0,\mathbf{x}\neq\mathbf{e}_i)||p((\mathbf{x},y)|\sigma_i<0,\mathbf{x}\neq\mathbf{e}_i)\right). \end{aligned}$$

Since the distribution of y is independent of σ_i, conditioned on $\mathbf{x}\neq\mathbf{e}_i$, this equals

$$\frac{1}{d'}D_{kl}\left(p(y|\sigma_i>0,\mathbf{x}=\mathbf{e}_i)||p(y|\sigma_i<0,\mathbf{x}=\mathbf{e}_i)\right). \tag{4}$$

The divergence in this equation is simply the KL divergence between two Bernoulli random variables, one with parameter $\frac{1}{2}(1+b)$, and the other with parameter $\frac{1}{2}(1-b)$. We now use Lemma 4 to upper bound (4) by

$$\frac{b^2}{d'}\left(\frac{1}{\frac{1}{2}(1+b)}+\frac{1}{\frac{1}{2}(1-b)}\right)=\frac{2b^2}{d'}\left(\frac{1}{1+b}+\frac{1}{1-b}\right)\leq\frac{2b^2}{d'}\left(1+\frac{1}{1/2}\right)=\frac{6b^2}{d'},$$

where we used the fact that $b\in[0,1/2]$. Summarizing the discussion so far, we showed that

$$D_{kl}\left(p(S|\sigma_i<0)||p(S|\sigma_i>0)\right) \ = \ m\ D_{kl}\left(p((\mathbf{x},y)|\sigma_i<0)||p((\mathbf{x},y)|\sigma_i>0)\right) \ = \ \frac{6mb^2}{d'}.$$

Plugging this back into (3), we get that the excess risk is lower bounded by

$$\begin{aligned} \frac{1}{2d'}\left(\min\{Yb,B/\sqrt{d'}\}\right)^2\sum_{i=1}^{d'}\left(1-\sqrt{\frac{3mb^2}{d'}}\right) &= \left(\min\{Yb,B/\sqrt{d'}\}\right)^2\frac{1}{2}\left(1-\sqrt{\frac{3mb^2}{d'}}\right) \\ &\geq \left(\min\{Yb,B/\sqrt{d'}\}\right)^2\frac{1}{2}\left(1-\sqrt{\frac{3m(d'/6m)}{d'}}\right) \\ &\geq 0.14\left(\min\{Yb,B/\sqrt{d'}\}\right)^2 \\ &= 0.14\left(\min\left\{Y\min\left\{\frac{1}{2},\sqrt{\frac{d'}{6m}}\right\},\frac{B}{\sqrt{d'}}\right\}\right)^2 \\ &= 0.14\min\left\{\frac{1}{4}Y^2,\frac{d'Y^2}{6m},\frac{B^2}{d'}\right\}. \end{aligned}$$

3. Specifically, we're using the fact that by Jensen's inequality, for any probability distributions p_1,p_2,q_1,q_2 and $\lambda\in[0,1]$, it holds that $D_{KL}((1-\lambda)p_1+\lambda p_2||(1-\lambda)q_1+\lambda q_2)\leq(1-\lambda)D_{KL}(p_1||q_1)+\lambda D_{KL}(p_2||q_2)$. See also Cover and Thomas (2006), theorem 2.7.2.

Now, recall that d' is a free parameter of value at most d. We now distinguish between two cases:

- If $d > \sqrt{6m}B/Y$, then we pick $d' = \lceil \sqrt{6m}B/Y \rceil$, and get that the expression above is at least

$$\begin{aligned} 0.14 \min\left\{\frac{1}{4}Y^2, \frac{B^2}{d'}\right\} &\geq 0.14 \min\left\{\frac{1}{4}Y^2, \frac{B^2}{\max\left\{1, 2\sqrt{6m}\frac{B}{Y}\right\}}\right\} \\ &= 0.14 \min\left\{\frac{1}{4}Y^2, B^2, \frac{BY}{2\sqrt{6m}}\right\}. \end{aligned}$$

- If $d \leq \sqrt{6m}B/Y$, we pick $d' = d$, and note that $\frac{d'Y^2}{6m} \leq \frac{B^2}{d}$ in this case. Therefore, the expression above is at least

$$0.14 \min\left\{\frac{1}{4}Y^2, \frac{dY^2}{6m}\right\}$$

Combining the two cases, we get that a lower bound of the form

$$c \ \min\left\{Y^2, B^2, \frac{dY^2}{m}, \frac{BY}{\sqrt{m}}\right\},$$

where c is a universal constant. ■

With Thm. 3 and Thm. 5 at hand, we now turn to prove our main result:

Proof [Proof of Thm. 1] Taking the maximum of Thm. 3 and Thm. 5, and using the fact that $B \geq 2Y$, we get a lower bound of

$$c \max\left\{\min\left\{Y^2, \frac{B^2}{m}\right\} \ , \ \min\left\{Y^2, \frac{dY^2}{m}, \frac{BY}{\sqrt{m}}\right\}\right\}$$

for some constant c. If $m \leq (B^2/Y^2)$, this is at least Y^2, and otherwise it is

$$\begin{aligned} c \max\left\{\frac{B^2}{m} \ , \ \min\left\{\frac{dY^2}{m}, \frac{BY}{\sqrt{m}}\right\}\right\} &\geq \frac{c}{2}\left(\frac{B^2}{m} + \min\left\{\frac{dY^2}{m}, \frac{BY}{\sqrt{m}}\right\}\right) \\ &\geq \frac{c}{2}\min\left\{\frac{B^2 + dY^2}{m}, \frac{BY}{\sqrt{m}}\right\}. \end{aligned}$$

Combining the two cases, the result follows. ■

Acknowledgments

We thank Nati Srebro and the anonymous reviewers for several helpful comments. This research is partially supported by an Israeli Science Foundation grant (no. 425/13) and an FP7 Marie Curie CIG grant.

References

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

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

K. Azoury and M. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006.

A. Gibbs and F. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

D. Hsu, S. M Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011.

G. Lecué and S. Mendelson. Performance of empirical risk minimization in linear aggregation. *arXiv Preprint arXiv:1402.5763*, 2014.

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

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

A. Singer, S. Kozat, and M. Feder. Universal linear least squares prediction: upper and lower bounds. *Information Theory, IEEE Transactions on*, 48(8):2354–2362, 2002.

N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2199–2207, 2010.

A. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer, 2003.

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.