Low ℓ_1-Norm and Guarantees on Sparsifiability

Shai Shalev-Shwartz and **Nathan Srebro**
Toyota Technological Institute—Chicago, USA
{shai,nati}@tti-c.org

June 26, 2008

Abstract

We consider the following problem: given a linear predictor $\mathbf{w}$ with low ℓ_1-norm, is it always possible to obtain a sparse predictor with similar error? It is interesting to understand this question as a further step in understanding the relationship between sparsity and the ℓ_1-norm, which is often used as a surrogate to sparsity. We show that for any $\epsilon > 0$, there exists a predictor with expected loss at most ϵ more than $\mathbf{w}$ that uses only $O\left(\left(\|\mathbf{w}\|_1/\epsilon\right)^2\right)$ features. Furthermore, such a predictor can be obtained using a simple randomized procedure. We show that this bound is tight, and hence the simple randomized procedure is in a sense optimal.

1 Introduction

Even when many features might be available for use in a prediction task, it is often beneficial to use only a small subset of the available features. Predictors that use only a small subset of features require a smaller memory footprint and can be applied faster. Furthermore, in applications such as medical diagnostics, obtaining each possible "feature" (e.g. test result) can be costly, and so a predictor that uses only a small number of features is desirable, even at the cost of a small degradation in performance relative to a predictor that uses more features. Focusing on linear prediction, it is generally difficult to find the best predictor subject to a constraint on the number of features used (the *sparsity* of the predictor), as this is a non-convex constraint that leads to a non-convex optimization problem. A common alternative is to seek a good predictor with small ℓ_1-norm, using this measure as a surrogate for sparsity. However, the resulting predictor need not necessarily be sparse. A common approach is to somehow obtain a sparse predictor from the learned low-ℓ_1-norm predictor. But can this always be done without significantly sacrificing performance?

In this paper we study the question of "sparsification" of linear predictors. Can a low-ℓ_1-norm predictor always be sparsified? I.e., does the existence of a good linear predictor with low ℓ_1-norm guarantee the existence of a good linear predictor that uses only a small number of features? If so, what is the relationship between the ℓ_1-norm and the number of features necessary to achieve similar performance? And is there a simple procedure for "sparsifying" a predictor, i.e. obtaining a good sparse predictor from a good predictor with low ℓ_1-norm?

We provide a simple randomized procedure for obtaining a sparse predictor $\tilde{\mathbf{w}}$ from a low-ℓ_1-predictor $\mathbf{w}$. We show that for any allowed degradation $\epsilon > 0$, our sparsification procedure can produce a linear predictor $\tilde{\mathbf{w}}$ that uses only $O\left(\left(\|\mathbf{w}\|_1/\epsilon\right)^2\right)$ features (independent of the overall number of features used by $\mathbf{w}$), and has expected loss at most ϵ worse than $\mathbf{w}$. Furthermore, we show that this relationship is tight (in the worst case): as many as $\Omega\left(\left(\|\mathbf{w}\|_1/\epsilon\right)^2\right)$ features might be required in order to get within ϵ of the expected loss of a linear predictor $\mathbf{w}$. We also show that the existence of a predictor with low ℓ_2-norm is *not* enough to guarantee the existence of a sparse predictor. This is perhaps not surprising, and provides further insight as to why ℓ_1-regularization is preferable to ℓ_2-regularization when sparsity is the true objective. Finally, we show that the common sparsification heuristic, in which the smallest elements of $\mathbf{w}$ are zeroed, might produce poor sparse predictors. For constructing our tightness results we derive a generalization of Khintchine inequality that holds for biased random variables. We believe that this inequality can be useful for deriving additional lower bounds in machine learning, involving linear loss functions.

Related work The use of the ℓ_1-norm as a surrogate for sparsity has a long history (e.g. [10] and the references therein), and much work has been done on understanding the relationship between the ℓ_1-norm and sparsity.

Donoho [2] provides sufficient conditions for when the optimal ℓ_1-norm predictor will be sparse, but this does not resolve the question of what happens when these conditions are not met and the optimal ℓ_1-norm predictor is *not* sparse, but we still desire a sparse classifier. Recent work on compressed sensing [1, 3] further explores how ℓ_1-norm regularization can be used for recovering a sparse predictor, but only under sever assumptions on the training examples (i.e. the design matrix).

Ng [6] considers PAC learning of a sparse predictor, and shows that ℓ_1-norm regularization is competitive with the best sparse predictor, while ℓ_2-regularization does not appear to be. In such a scenario we are not interested in the resulting predictor being sparse (it won't necessarily be sparse), but only in its generalization performance. In contrast, in this paper we *are* interested in the resulting predictor being sparse, but do not study ℓ_1-regularized learning. Rather, we assume we already have a good low-ℓ_1-norm predictor, and ask whether we can obtain from it a good predictor that is sparse.

The converse of our question, focusing on linear classification, was recently resolved by Servedio [9]: given a sparse linear separator, can it always be represented using small weights?

The randomized sparsification procedure we suggest was previously proposed by Schapire et al [8], as a tool for obtaining generalization bounds for boosting. However, Schapire et al's bound depends on $\log(m)$, where m is the number of examples in the input distribution, and is therefore only valid for guaranteeing performance over a finite sample. Our bound does not depend on m and is adequate for guaranteeing performance over an arbitrary source distribution.

Studying neural networks with bounded fan-in, Lee et al [5] addressed an equivalent formulation of this question, providing an upper bound similar to ours, for the special case of the squared-error loss. Here we obtain a more general result, that holds for any (Lipschitz-continuous) loss function. Furthermore, we present matching upper

and lower bounds, which together tightly characterize the possible sparseness guaranteed by low ℓ_1-norm.

2 Problem Setting

We first introduce our notation and formally describe the problem setting. We denote scalars with lower case letters (e.g. x) and vectors with bold-face letters (e.g. $\mathbf{x}$). Random variables are designated by sans-serif fonts (e.g. x) and random vectors by bold-face sans-serif fonts (e.g. $\mathbf{x}$). The set of non-negative reals is denoted by $\mathbb{R}_+$, and the set of integers $\{1, \ldots, k\}$ is denoted by $[k]$.

Consider instances represented by vectors of n features $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$ and target values y in some target space $\mathcal{Y}$. E.g., in classification $\mathcal{Y} = \{+1, -1\}$, while in regression $\mathcal{Y} = \mathbb{R}$. A linear predictor is a function of the form[1] $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle = \sum_{i=1}^{n} w_i x_i$. Given an instance-target pair $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$, we assess the quality of a predictor $\mathbf{w}$ for this pair using a loss function $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$. For example, a well studied loss function for classification is the hinge-loss defined as $L(\langle \mathbf{w}, \mathbf{x} \rangle, y) = \max\{0, 1 - y \langle \mathbf{w}, \mathbf{x} \rangle\}$. We say that a loss function is λ-Lipschitz with respect to its first argument if for any two scalars $a_1, a_2 \in \mathbb{R}$ and target value $y \in Y$ we have that $|L(a_1, y) - L(a_2, y)| \leq \lambda |a_1 - a_2|$. For a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, we assess the quality of the predictor $\mathbf{w}$ using its expected loss $\mathbb{E}_{(\mathbf{x},\mathsf{y})\sim\mathcal{D}}[L(\langle \mathbf{w}, \mathbf{x} \rangle, \mathsf{y})]$.

The focus of this paper is on constructions of sparse linear predictors. We refer to the number of non-zero elements of a predictor $\mathbf{w}$ as its "sparsity", and denote it $\|\mathbf{w}\|_0$. Predictor *sparsification* is a procedure which receives as input an arbitrary (possibly non-sparse) predictor $\mathbf{w}$ and a target sparsity level S and returns as output a predictor $\tilde{\mathbf{w}}$ for which $\|\tilde{\mathbf{w}}\|_0 \leq S$. Naturally, the sparsification procedure might damage the accuracy of the predictor $\mathbf{w}$. That is, the expected loss of the sparse predictor $\tilde{\mathbf{w}}$ might be larger than the expected loss of the non-sparse predictor $\mathbf{w}$. The main result of this paper is a sparsification procedure along with sufficient conditions under which the excess loss of the resulting sparse predictor is small.

3 Guaranteed Sparsification Procedure

We are now ready to present our randomized "sparsification" procedure. That is, a procedure that takes as input a low ℓ_1-norm predictor and outputs a sparse predictor with similar performance. We then state our main theorem which bounds the degradation in performance of the sparsified predictor as a function of the ℓ_1-norm of the original predictor and the desired sparsity.

Let $\mathbf{w} \in \mathbb{R}^n$ be an arbitrary (possibly dense) predictor. Without loss of generality, we assume that $w_j \geq 0$ for all j (since otherwise, if $w_j < 0$, we can flip the sign of the j'th feature). Thus, the predictor $\mathbf{w}/\|\mathbf{w}\|_1$ defines a probability measure over the set $[n]$. To motivate our construction we would like to note that the prediction $\langle \mathbf{w}, \mathbf{x} \rangle$ can be viewed as the expected value of the elements in $\mathbf{x}$ according to the

[1] For clarity of presentation, in this paper we do not allow an unregularized biased term. Both our sparsifiability results and tightness results can be easily modified to allow a bias term

distribution $\mathbf{w}/\|\mathbf{w}\|_1$ (scaled by $\|\mathbf{w}\|_1$). We can approximate this expected value by an empirical average of randomly selected elements of $\mathbf{x}$. Since our goal is to find a sparse predictor whose predictions are similar to those of $\mathbf{w}$, we construct the sparse predictor by randomly selecting S elements from $[n]$ based on the probability measure $\mathbf{w}/\|\mathbf{w}\|_1$.

Formally, let $\mathbf{r} = (\mathsf{r}_1, \ldots, \mathsf{r}_S)$ be a sequence of i.i.d. random variables over $[n]$ where for all $i \in [S]$ and $j \in [n]$ we have $\mathbb{P}(\mathsf{r}_i = j) = \frac{w_j}{\|\mathbf{w}\|_1}$. Let $\mathbf{e}^i \in \mathbb{R}^n$ be the vector whose jth element is zero if $i \neq j$ and 1 if $i = j$. We set our sparse predictor to be

$$\tilde{\mathbf{w}} \;=\; \frac{\|\mathbf{w}\|_1}{S} \sum_{i=1}^{S} \mathbf{e}^{\mathsf{r}_i} \;\; . \tag{1}$$

For any given $\mathbf{x} \in \mathcal{X}$ we have,

$$\mathbb{E}[\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle] \;\; = \;\; \frac{\|\mathbf{w}\|_1}{S} \sum_{i=1}^{S} \mathbb{E}[x_{\mathsf{r}_i}] \;=\; \|\mathbf{w}\|_1 \, \mathbb{E}[x_{\mathsf{r}_1}] \;=\; \|\mathbf{w}\|_1 \sum_{j=1}^{n} \frac{w_j}{\|\mathbf{w}\|_1} x_j \;=\; \langle \mathbf{w}, \mathbf{x} \rangle \;\; .$$

Thus, the expectations of the predictions of $\tilde{\mathbf{w}}$ are precisely the predictions of $\mathbf{w}$ (where the expectation is over the sparsification). The theorem below states that if $\|\mathbf{w}\|_1$ is significantly smaller than $\sqrt{S}$ then the predictions of $\tilde{\mathbf{w}}$ are concentrated around their expectation, and are therefore very similar to those of $\mathbf{w}$, yielding almost the same performance.

Theorem 1 *Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_\infty \leq 1\}$ be an instance space, $\mathcal{Y}$ be a target space, $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$ and $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a loss function which is λ-Lipschitz with respect to its first argument. For any $\mathbf{w} \in \mathbb{R}^n_+$, let $\boldsymbol{\mathsf{r}} = (\mathsf{r}_1, \ldots, \mathsf{r}_S)$ be a sequence of independent random variables over $[n]$, each of which is distributed according to $\mathbf{w}/\|\mathbf{w}\|_1$, and let $\tilde{\boldsymbol{\mathsf{w}}}$ be the random predictor defined in Eq. 1. Then, for any scalar $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the choice of $\boldsymbol{\mathsf{r}}$ we have that*

$$\mathbb{E}_{(\boldsymbol{\mathsf{x}},\mathsf{y})\sim\mathcal{D}}[L(\langle \tilde{\boldsymbol{\mathsf{w}}}, \boldsymbol{\mathsf{x}} \rangle, \mathsf{y})] \;\leq\; \mathbb{E}_{(\boldsymbol{\mathsf{x}},\mathsf{y})\sim\mathcal{D}}[L(\langle \mathbf{w}, \boldsymbol{\mathsf{x}} \rangle, \mathsf{y})] + \sqrt{2}\, \frac{\lambda \|\mathbf{w}\|_1}{\sqrt{S}} \left(\sqrt{\log(1/\delta)} + 5 \right) \;\; .$$

Before proving Theorem 1, we underscore its consequences. First, note that in order to establish the existence of a sparse predictor, we could take $\delta \to 1$. We get that the existence of a predictor $\mathbf{w}$ with expected loss l, guarantees the existence of a sparse predictor $\tilde{\mathbf{w}}$, with $\|\tilde{\mathbf{w}}\|_0 \leq (7.1\, \lambda \, \|\mathbf{w}\|_1 / \epsilon)^2$ and expected loss at most $l + \epsilon$. Furthermore, a good sparse predictor can be obtained based on the knowledge of a good low-ℓ_1-norm predictor. Suppose we learned a predictor $\mathbf{w}$ with low generalization error (expected loss on the source distribution), $\mathbb{E}_{(\mathbf{x},\mathsf{y})\sim\mathcal{D}}[L(\langle \mathbf{w}, \mathbf{x} \rangle, \mathsf{y})]$, where $\mathcal{D}$ is an (unknown) source distribution over $\mathcal{X} \times \mathcal{Y}$. We would now like to construct a sparse predictor, $\tilde{\mathbf{w}}$, such that the generalization error of $\tilde{\mathbf{w}}$ does not exceed that of $\mathbf{w}$ by more than ϵ. Theorem 1 tells us that if we choose $\tilde{\mathbf{w}}$ randomly as in Eq. 1 and if $S \geq (9 \|\mathbf{w}\|_1 / \epsilon)^2$, then there's a 99% chance that the generalization error of $\tilde{\mathbf{w}}$ does not exceed that of $\mathbf{w}$ by more than ϵ. Note that to perform the sparsification we do not

need access to the source distribution $\mathcal{D}$ nor to any samples—the sparse predictor $\tilde{\mathbf{w}}$ is a (random) function of only the (dense) predictor $\mathbf{w}$.

We can also use Theorem 1 in order to always find a sparse predictor with good performance on a given training set of examples, $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, given a low ℓ_1-norm predictor $\mathbf{w}$ with good training performance (e.g. one that was learned using this training set). To do so, construct a sparse predictor as in Eq. 1 and evaluate its performance on the training set of examples. Then, repeat the randomized construction until the average loss of $\tilde{\mathbf{w}}$ on the training set is at most ϵ plus the average loss of $\mathbf{w}$ on the training set. Setting $\mathcal{D}$ to be the uniform distribution over the training examples and applying Theorem 1 with $\delta = 0.5$ we get that with probability at least half, the performance of $\tilde{\mathbf{w}}$ does not exceed that of $\mathbf{w}$ by more than $8.3\,\lambda\,\|\mathbf{w}\|_1/\sqrt{S}$. Thus, if $S \geq (8.3\,\lambda\,\|\mathbf{w}\|_1/\epsilon)^2$, after an average of two random sparsification attempts, we will obtain an S-sparse predictor with mean training loss at most ϵ more than that of our original (dense) predictor $\mathbf{w}$.

Proof of Theorem 1 Recall that $\tilde{\mathbf{w}}$ is a function of the sequence of random variables $\mathbf{r} = (\mathsf{r}_1, \ldots, \mathsf{r}_S)$. Let us denote by $g(\mathbf{r})$ the expected loss of $\tilde{\mathbf{w}}$, namely,

$$g(\mathbf{r}) \;=\; \mathbb{E}_{(\mathbf{x},\mathsf{y})\sim D}[L(\langle \tilde{\mathbf{w}}, \mathbf{x}\rangle, \mathsf{y})] \;. \tag{2}$$

We prove the theorem using the following three steps. First, we use McDiarmid inequality to show that $g(\mathbf{r})$ is concentrated. Specifically, Lemma 1 states that with probability of at least $1-\delta$ we have

$$g(\mathbf{r}) \;\leq\; \mathbb{E}_{\mathbf{r}}[g(\mathbf{r})] + \frac{\lambda\,\|\mathbf{w}\|_1}{\sqrt{S}}\,\sqrt{2\log(2/\delta)}\,. \tag{3}$$

Next, in Lemma 2 we utilize the fact that L is λ-Lipschitz to show that

$$\mathbb{E}_{\mathbf{r}}[g(\mathbf{r})] \;\leq\; \mathbb{E}_{(\mathbf{x},\mathsf{y})}[L(\langle \mathbf{w}, \mathbf{x}\rangle, \mathsf{y})] + \lambda\,\mathbb{E}_{(\mathbf{x},\mathsf{y})}\mathbb{E}_{\mathbf{r}}[|\,\langle \tilde{\mathbf{w}}, \mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\,|]\,. \tag{4}$$

Finally, Lemma 3 bounds the expected difference between $\langle \mathbf{w}, \mathbf{x}\rangle$ and $\langle \tilde{\mathbf{w}}, \mathbf{x}\rangle$. This bound holds for *any* $\mathbf{x} \in \mathcal{X}$ and therefore also for the expectation over $\mathbf{x}$:

$$\mathbb{E}_{(\mathbf{x},\mathsf{y})}\mathbb{E}_{\mathbf{r}}[|\,\langle \tilde{\mathbf{w}}, \mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\,|] \;\leq\; 4\sqrt{2}\,\|\mathbf{w}\|_1/\sqrt{S}\,. \tag{5}$$

Combining Eqs. 3-5 gives that

$$g(\mathbf{r}) \;\leq\; \mathbb{E}_{(\mathbf{x},\mathsf{y})}[L(\langle \mathbf{w}, \mathbf{x}\rangle, \mathsf{y})] + \tfrac{\sqrt{2}\lambda\,\|\mathbf{w}\|_1}{\sqrt{S}}\,\left(\sqrt{\log(2/\delta)} + 4\right)\,.$$

The inequality in Theorem 1 follows from the above by further bounding $\sqrt{\log(2/\delta)} \leq \sqrt{\log(1/\delta)} + \sqrt{\log(2)}$ and $4 + \sqrt{\log(2)} < 5$. □

Lemma 1 *Under the conditions of Theorem 1, let $g(\boldsymbol{r})$ be as defined in Eq. 2 and let $\delta \in (0,1)$. Then, the following bound holds with probability of at least $1-\delta$,*

$$g(\boldsymbol{r}) \;\leq\; \mathbb{E}_{\boldsymbol{r}}[g(\boldsymbol{r})] + \frac{\lambda\,\|\mathbf{w}\|_1}{\sqrt{S}}\,\sqrt{2\log(2/\delta)}\,.$$

Proof We prove the lemma using McDiarmid inequality. First, we need to show that g has the bounded differences property with parameter $\lambda \|\mathbf{w}\|_1/S$. That is, we need to show that for all $i \in [S]$ and $r'_i \in [n]$, the difference between $g(\mathsf{r}_1, \ldots, \mathsf{r}_S)$ and $g(\mathsf{r}_1, \ldots, \mathsf{r}_{i-1}, r'_i, \mathsf{r}_{i+1}, \ldots, \mathsf{r}_S)$ is at most $2\lambda \|\mathbf{w}\|_1/S$. To do so, we denote by $\tilde{\mathbf{w}}'$ the vector $\tilde{\mathbf{w}}' = \frac{\|\mathbf{w}\|_1}{S}(\mathbf{e}^{r'_i} + \sum_{j \neq i} \mathbf{e}^{\mathsf{r}_i})$. Note that for all $\mathbf{x} \in \mathcal{X}$ we have that, $|\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle - \langle \tilde{\mathbf{w}}', \mathbf{x} \rangle| = \frac{\|\mathbf{w}\|_1}{S} \left| x_{\mathsf{r}_i} - x_{r'_i} \right| \leq \frac{2\|\mathbf{w}\|_1}{S}$. Since L is λ-Lipschitz the above implies that $|L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle, \mathrm{y}) - L(\langle \tilde{\mathbf{w}}', \mathbf{x} \rangle, \mathrm{y})| \leq \frac{2\lambda\|\mathbf{w}\|_1}{S}$. The last inequality holds for any $\mathbf{x} \in \mathcal{X}$ and therefore we conclude that

$$|\, \mathbb{E}_{(\mathbf{x},\mathrm{y})}[L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle, \mathrm{y})] - \mathbb{E}_{(\mathbf{x},\mathrm{y})}[L(\langle \tilde{\mathbf{w}}', \mathbf{x} \rangle, \mathrm{y})]\,| \;\leq\; 2\lambda \|\mathbf{w}\|_1/S\,.$$

We have thus shown that g has the bounded differences property. We can now utilize McDiarmid inequality to get that for all $t > 0$ we have that $\mathbb{P}[|g - \mathbb{E}[g]| \geq t] \leq 2\exp\left(-St^2/(2\lambda^2 \|\mathbf{w}\|_1{}^2)\right)$. Denote by δ the right-hand side of the above and solving for t we conclude that with probability of at least $1-\delta$ the following holds, $g(\mathbf{r}) \leq \mathbb{E}[g(\mathbf{r})] + \frac{\lambda\|\mathbf{w}\|_1}{\sqrt{S}}\sqrt{2\log(2/\delta)}$. □

Lemma 2 *Under the conditions of Lemma 1, we have*

$$\mathbb{E}_{\boldsymbol{r}}[g(\boldsymbol{r})] \;\leq\; \mathbb{E}_{(\boldsymbol{x},\mathrm{y})}[L(\langle \mathbf{w}, \boldsymbol{x} \rangle, \mathrm{y})] + \lambda\, \mathbb{E}_{(\boldsymbol{x},\mathrm{y})}\mathbb{E}_{\boldsymbol{r}}[|\, \langle \tilde{\boldsymbol{w}}, \boldsymbol{x} \rangle - \langle \mathbf{w}, \boldsymbol{x} \rangle\,|]\,.$$

Proof Fix $\mathbf{x} \in \mathcal{X}$. Since L is λ-Lipschitz we have $L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle, \mathrm{y}) \leq L(\langle \mathbf{w}, \mathbf{x} \rangle, \mathrm{y}) + \lambda\,|\, \langle \tilde{\mathbf{w}}, \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle\,|$. Taking expectation of the above over $\mathbf{r}$:

$$\mathbb{E}_{\mathbf{r}}[L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle), \mathrm{y}] \leq L(\langle \mathbf{w}, \mathbf{x} \rangle, \mathrm{y}) + \lambda\, \mathbb{E}_{\mathbf{r}}[|\, \langle \tilde{\mathbf{w}}, \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle\,|]\,.$$

Taking again expectation of the above this time over $(\mathbf{x}, \mathrm{y})$ gives,

$$\mathbb{E}_{(\mathbf{x},\mathrm{y})}\mathbb{E}_{\mathbf{r}}[L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle), \mathrm{y}] \leq\; \mathbb{E}_{(\mathbf{x},\mathrm{y})}[L(\langle \mathbf{w}, \mathbf{x} \rangle, \mathrm{y})] + \lambda\, \mathbb{E}_{(\mathbf{x},\mathrm{y})}\mathbb{E}_{\mathbf{r}}[|\, \langle \tilde{\mathbf{w}}, \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle\,|]\,.$$

The left-hand side of the above equation equals to $\mathbb{E}_{\mathbf{r}}\mathbb{E}_{(\mathbf{x},\mathrm{y})}[L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle), \mathrm{y}]$, which by definition equals to $\mathbb{E}_{\mathbf{r}}[g(\mathbf{r})]$ and our proof is concluded. □

Lemma 3 *Under the conditions of Theorem 1 we have that for all* $\mathbf{x} \in \mathcal{X}$*,*

$$\mathbb{E}_{\boldsymbol{r}}\,[\,|\, \langle \tilde{\boldsymbol{w}}, \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle\,|\,] \leq 4\sqrt{2}\,\tfrac{\|\mathbf{w}\|_1}{\sqrt{S}}\,.$$

Proof For all $i \in [n]$ denote $z_i = \|\mathbf{w}\|_1\, x_{\mathsf{r}_i}$. The sequence $z_1, \ldots, z_S$ is an independently and identically distributed sequence with $|z_i| \leq \|\mathbf{w}\|_1$ for all $i \in [n]$ since $\|\mathbf{x}\|_\infty \leq 1$. We can rewrite $\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle$ as $\sum_{i=1}^S \frac{\|\mathbf{w}\|_1}{S}\, x_{\mathsf{r}_i} = \frac{1}{S}\sum_{i=1}^S z_i$. Therefore, $\mathbb{E}[\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle] = \langle \mathbf{w}, \mathbf{x} \rangle$. Using Hoeffding inequality we get that for $t > 0$, $\mathbb{P}[\,|\, \langle \tilde{\mathbf{w}}, \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle\,| > t\,] \leq 2\exp\left(-t^2 S/(2\|\mathbf{w}\|_1^2)\right)$. The last inequality tells us that the random variable $|\, \langle \tilde{\mathbf{w}}, \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle\,|$ is concentrated around zero. Lemma 5 in Appendix A states that the expectation of a random variable that is concentrated around zero cannot be very large. Formally, applying Lemma 5 with $a = \|\mathbf{w}\|_1\sqrt{2/S}$ concludes our proof. □

4 Tightness and Extreme Examples

In the previous section, we established that the existence of a predictor $\mathbf{w}$ with $\|\mathbf{w}\|_1 = B$ guarantees the existence of a sparse predictor $\tilde{\mathbf{w}}$ with $\|\tilde{\mathbf{w}}\|_0 = O(B^2/\epsilon^2)$ and expected loss at most ϵ more than the expected loss of w. We further saw how such a sparse predictor can be obtained. We now argue that this relationship is tight, and a better guarantee cannot be obtained. The procedure of the previous section is therefore optimal in the sense that no other procedure can yield a better sparsity guarantee (better by more than a constant factor) in terms of the ℓ_1-norm of the input predictor w.

We will use the following lemma (proved in Appendix A), which generalizes the Khintchine inequality also to biased random variables. We use the lemma in order to obtain lower bounds on the mean-absolute error in terms of the bias and variance of the prediction:

Lemma 4 *Let $\boldsymbol{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_n)$ be a sequence of independent Bernoulli random variables with $0.05 \leq \mathbb{P}[\mathsf{x}_k = 1] \leq 0.95$. Let Q be an arbitrary polynomial over n variables of degree d. Then,*

$$\mathbb{E}[\,|Q(\boldsymbol{x})|\,] \;\geq\; (0.2)^d\, \mathbb{E}[\,|Q(\boldsymbol{x})|^2\,]^{\frac{1}{2}} \;\;.$$

Theorem 2 *For any $B > 2$ and $l > 0$, there exists a data distribution, such that a (dense) predictor $\mathbf{w}$ with $\|\mathbf{w}\|_1 = B$ can achieve mean absolute-error ($L(a,b) = |a-b|$) less than l, but for any $\epsilon \leq 0.1$, at least $B^2/(45\,\epsilon^2)$ features must be used for achieving mean absolute-error less than ϵ.*

Proof Fix some $B > 2$, $l > 0$, and $\epsilon < 0.1$. To prove the theorem, we present an input distribution $\mathcal{D}$, then demonstrate a specific (dense) predictor with $\|\mathbf{w}\|_1 = B$ and mean absolute-error l, and finally present a lower bound on mean absolute-error of any sparse predictor, from which we can conclude that any predictor $\mathbf{u}$ with mean-absolute error at most ϵ must satisfy $\|\mathbf{u}\|_0 \geq B^2/(45\,\epsilon^2)$.

The data distribution: Consider an instance space $\mathcal{X} = \{+1, -1\}^n$, where $n \geq 1/(la)^2$, and a target space $\mathcal{Y} = \{+1, -1\}$. The distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is as follows. First, the label y is uniformly distributed with $\mathbb{P}(\mathsf{y} = 1) = \frac{1}{2}$. Next, the features $\mathsf{x}_1, \ldots, \mathsf{x}_n$ are identically distributed and are independent conditioned on y, with $\mathbb{P}(\mathsf{x}_i = \mathsf{y} \mid \mathsf{y}) = \frac{1+a}{2}$, where $a = 1/B$. Thus, the correlation between each feature and the label is $1/B$. In such an example, the "information" about the label is spread among all features, and in order to obtain a good predictor, this distributed information needs to be pulled together, e.g. using a dense linear predictor.

A dense predictor: Consider the predictor $\mathbf{w}$ with $w_i = 1/(na)$ for all features i. To simplify our notation, we use the shorthand $\mathbb{E}[\,\langle \mathbf{w}, \mathbf{x}\rangle \mid y]$ for denoting $\mathbb{E}[\,\langle \mathbf{w}, \mathbf{x}\rangle \mid \mathsf{y} = y]$. Verifying that $\mathbb{E}[\,\langle \mathbf{w}, \mathbf{x}\rangle \mid y] = n\frac{1}{na}\, a\, y = y$ for both values of y, and using Jensen's inequality we obtain that:

$$\mathbb{E}[\,|\langle \mathbf{w}, \mathbf{x}\rangle - \mathsf{y}| \mid y] = \mathbb{E}[\,|\langle \mathbf{w}, \mathbf{x}\rangle - \mathbb{E}[\,\langle \mathbf{w}, \mathbf{x}\rangle \mid y]| \mid y] \tag{6}$$

$$\leq \sqrt{\mathbb{E}\left[\,(\langle \mathbf{w}, \mathbf{x}\rangle - \mathbb{E}[\,\langle \mathbf{w}, \mathbf{x}\rangle \mid y])^2 \;\middle|\; y\right]} = \sqrt{\mathrm{Var}[\,\langle \mathbf{w}, \mathbf{x}\rangle \mid y]} = \sqrt{\tfrac{1-a^2}{na^2}} \leq \tfrac{1}{\sqrt{n}a} \leq l \;\;.$$

Since the expected loss is bounded by l for both values of y, we conclude that $\mathbb{E}[|\langle \mathbf{w}, \mathbf{x} \rangle - \mathsf{y}|] \leq l$. In summary, we get a predictor with ℓ_1-norm $\|\mathbf{w}\|_1 = 1/a = B$ whose expected error is at most l.

Sparse prediction: Consider any predictor $\mathbf{u}$ with only S non-zero coefficients. For such a predictor we have $\sum \mathbf{u}_i^2 \geq (\sum \mathbf{u}_i)^2/S$. Since we consider only $B > 2$, we have $0.05 < 0.25 \leq \mathbb{P}[\mathsf{x}_i = \mathsf{y}|y] \leq 0.75 < 0.95$, with the loss being an affine function (degree one polynomial) of $\mathbf{x}$. We can therefore use Lemma 4 to get that:

$$\mathbb{E}\left[\,|\,\langle \mathbf{u}, \mathbf{x} \rangle - \mathsf{y}|\,|\,y\right] \;\geq\; 0.2\,\sqrt{\mathbb{E}\left[\,(\langle \mathbf{u}, \mathbf{x} \rangle - \mathsf{y})^2\,|\,y\right]} \;=\; 0.2\,\sqrt{\mathrm{Var}[\langle \mathbf{u}, \mathbf{x} \rangle\,|\,y] + (\mathbb{E}\left[\,\langle \mathbf{u}, \mathbf{x} \rangle\,|\,y\right] - y)^2}\,.$$

Next, we can calculate $\mathrm{Var}[\langle \mathbf{u}, \mathbf{x} \rangle\,|\,y] = (1-a^2)\sum_i u_i^2 \geq (1-a^2)(\sum_i u_i)^2/S$ and $\mathbb{E}\left[\,\langle \mathbf{u}, \mathbf{x} \rangle\,|\,y\right] = ya\sum_i u_i$. Denote $\rho = \sum_i u_i$ we therefore get that

$$\mathbb{E}[|\,\langle \mathbf{u}, \mathbf{x} \rangle - \mathsf{y}|] \;=\; \mathbb{E}_{\mathsf{y}}\mathbb{E}\left[\,|\,\langle \mathbf{u}, \mathbf{x} \rangle - \mathsf{y}|\,|\,\mathsf{y}\right] \;\geq\; 0.2\,\sqrt{(1-a^2)\rho^2/S + (a\rho - 1)^2}\,.$$

The argument inside the square-root is a quadratic expression in ρ that achieves its minimum at $\rho^* = a/(\frac{1-a^2}{S} + a^2)$. Substituting $\rho = \rho^*$ we can bound the expected loss by:

$$\mathbb{E}[|\,\langle \mathbf{u}, \mathbf{x} \rangle - \mathsf{y}|] \;\geq\; 0.2\,/\sqrt{1 + \tfrac{a^2}{1-a^2}S}\;\;. \tag{7}$$

Recalling that $a = 1/B < 1/2$, to get an expected loss of at most ϵ, for $\epsilon < 0.1$, we must have:

$$S \;\geq\; (B^2-1)\left(\tfrac{0.2^2-\epsilon^2}{\epsilon^2}\right) \;\geq\; \tfrac{B^2}{45\,\epsilon^2}\;\;.$$

We have shown that any predictor with $\mathbb{E}[|\langle \mathbf{u}, \mathbf{x} \rangle - \mathsf{y}|] \leq \epsilon$ must satisfy $\|\mathbf{u}\|_0 = S \geq B^2/(45\,\epsilon^2)$. □

Note that without using Lemma 4 the above arguments can be used to obtain a lower bound of $\|\mathbf{u}\|_0 = \Omega(B^2/\epsilon)$ on the sparsity of a predictor achieving squared-error at most ϵ (note the linear rather than squared dependence on ϵ). This lower bound for the special case of the squared error is tight and matches the upper-bound analysis of Lee et al [5].

4.1 Low ℓ_2-norm does not guarantee sparsifiability

One might ask if the existence of a predictor with low ℓ_2-norm can also guarantee the existence of a sparse predictor. Perhaps even if our proposed sparsification procedure does not work well on predictors with low ℓ_2-norm, a different procedure might be used to sparsify such predictors. We now show that this is not the case, by presenting examples where good predictions can be obtained by predictors with arbitrarily low ℓ_2-norm, but for which an arbitrarily high number of features is required in order to achieve a fixed performance.

To do so, we use the same type of data distribution and dense predictor as in the previous section. Setting $w_i = 1/(na)$ yields $\|\mathbf{w}\|_2 = 1/(a\sqrt{n})$. Therefore, we can decrease the correlation a as we increase the dimension n, keeping the ℓ_2-norm of the dense predictor fixed, but requiring an increasing number of features in order to obtain good performance.

Specifically, let B and l be arbitrarily small positive numbers and S arbitrarily large. Consider again the data distribution $\mathbb{P}(\mathsf{y} = 1) = 1/2$ and $\mathbb{P}(\mathsf{x}_i = y|y) = (1 + a)/2$ with a correlation of $a = (S/3 + 1)^{-1}$ and dimensionality $n = 1/(a \cdot \min(B, l))^2$. Following the calculations above (Eq. 6), the dense predictor with $w_i = \frac{1}{na}$, and so $\|\mathbf{w}\| = \frac{1}{a\sqrt{n}} \leq B$, achieves expected absolute-error $\mathbb{E}\left[|\langle \mathbf{w}, \mathbf{x}\rangle - \mathsf{y}|\right] \leq 1/(\sqrt{n}\, a) \leq l$. In contrast, using Eq. 7 we get that no predictor that uses less than S features can achieve expected absolute-error less than 0.1.

4.2 Sparsifying by considering only large weights

The procedure described in Section 3 involves random sampling of the features. An alternative deterministic procedure, commonly used in practice, is to choose only the features with the largest weights, or in other words, to zero small weights of the predictor (and perhaps readjust the remaining weights). We will consider applying this deterministic procedure to a low ℓ_1-norm predictor $\mathbf{w}^*$ learned by minimizing the expected loss subject to ℓ_1-norm regularization. Even on such an "optimal" predictor $\mathbf{w}^*$, using only the features with largest coefficients, can yield a large degradation in performance. This can happen when many features are highly correlated.

Specifically, for any arbitrarily large S and arbitrarily small l, we show an example in which the optimal predictor $\mathbf{w}^*$ with ℓ_1-norm at most 3 achieves mean absolute-error at most l, but using any re-weighting of the S features with the largest coefficients yields mean absolute-error of at least 0.02. Note that if $S \geq (25/\epsilon)^2$, our randomized procedure would yield a S-sparse predictor with error at most $l + \epsilon$, which we could set arbitrarily close to zero.

We again define a joint distribution over binary targets $\mathsf{y} \in \{+1, -1\}$ and binary feature vectors $\mathbf{x} \in \{+1, -1\}^{Sn}$, with $n = 7/l^2$ (i.e. the overall dimensionality is $7S/l^2$). For convenience we will label the features with two indices: $\mathsf{x}_{1,1}, \ldots, \mathsf{x}_{S,n}$. To describe the data distribution we use another set of n (latent) binary random variables $\mathsf{z}_1, \ldots, \mathsf{z}_n \in \{+1, -1\}$, i.i.d. given y, with $\mathbb{P}[\mathsf{z}_i = \mathsf{y} \,|\, \mathsf{y}] = (1 + \frac{1}{3} + \frac{i-1}{3(n-1)})/2$ (i.e. the correlation between these variables and the labels are between $1/3$ and $2/3$). The features $\mathsf{x}_{i,j}$ are independent given $\mathbf{z}, \mathsf{y}$, and are specified by $\mathbb{P}[\mathsf{x}_{i,j} = \mathsf{z}_i \,|\, \mathsf{z}_i] = 7/8 = (1 + 0.75)/2$. The features are thus grouped into n groups of S highly correlated features, where the correlation between each feature and the label varies between $1/4$ and $1/2$.

The minimum-mean-absolute-error predictor among those with ℓ_1-norm bounded by three, $\mathbf{w}^* = \arg\min_{\|\mathbf{w}\|_1 \leq 3} \mathbb{E}[|\, \langle \mathbf{w}, \mathbf{x}\rangle - \mathsf{y}|]$, achieves mean absolute-error less than l (Appendix B). However, in this optimal predictor, the weights $w_{n,i}$ corresponding to features in the last group will be larger than any other weights (Appendix B), and so these features will be selected as the maximal weight features. But since these features are all highly correlated, using any combination of them will not yield mean absolute-error better than 0.02 (Appendix B).

We also note that the features in the last group would also be the first S features selected by following the ℓ_1-norm regularization path or by related methods such as LARS [4].

5 Discussion

Using the ℓ_1-norm as a surrogate to sparsity is prevalent in machine learning and other domains. It is therefore interesting to precisely understand the relationship between the ℓ_1-norm and sparsity. Here we answer a fundamental question in this regard: how much sparsity does low ℓ_1-norm guarantee? This question is relevant when we are directly interested in obtaining a predictor that uses only a small number of features, e.g. when each feature is expensive to compute. We show here that indeed having low ℓ_1-norm does guarantee the existence of a sparse predictor, and present a simple randomized procedure for obtaining such a sparse predictor. We also precisely characterize what level of sparsity one can hope for, and show that the randomized procedure does indeed achieve this optimal sparsity (up to a constant factor).

We emphasize that our results assume we already have a low ℓ_1-norm predictor in hand, and indeed our randomized procedure depends only on this predictor, and not on the data. We do not address here the complimentary question of the appropriateness of using ℓ_1-norm regularization when we would like to be competitive with some unknown sparse predictor. This question has been recently addressed by Ng [6], who argued that in this PAC-learning setup the ℓ_1-norm is much more appropriate than the ℓ_2-norm as a surrogate to sparsity. Here, we see from another perspective how the ℓ_1-norm, and not the ℓ_2-norm, is much more closely related to sparsity.

References

[1] E. J. Candes. Compressive sampling. *Proc. of the Int. Congress of Math., Madrid, Spain*, 2006.

[2] D.L. Donoho. For most large underdetermined systems of linear equations, the minimal l1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math. 59*, 2006.

[3] D.L. Donoho. Compressed Sensing. Technical Report, Stanford University, 2006.

[4] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani. Least Angle Regression. *Annals of Statistics*, 32(2), 2004.

[5] W.S. Lee, P.L. Bartlett, and R.C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

[6] A.Y. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *ICML*, 2004.

[7] K. Oleszkiewicz. On a nonsymmetric version of the Khinchine-Kahane inequality. *Progress In Probability*, 56:156, 2003.

[8] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics, 26(5)*, 1998.

[9] R. Servedio. Every linear threshold function has a low-weight approximator. In *18th Annual Conf. on Comp. Complexity (CCC)*, 2006.

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

A Additional Lemmas

Proof of Lemma 4 Using Holder's inequality with $p = 3/2$ and $q = 3$ we have

$$\begin{aligned}\mathbb{E}[\,|Q(\mathbf{x})|^2\,] &= \sum_{\mathbf{x}\in\{0,1\}^n} \mathbb{P}(\mathbf{x})|Q(\mathbf{x})|^2 = \sum_{\mathbf{x}} \left(\mathbb{P}(\mathbf{x})^{2/3}|Q(\mathbf{x})|^{2/3}\right)\left(\mathbb{P}(\mathbf{x})^{1/3}|Q(\mathbf{x})|^{4/3}\right) \\ &\leq (\sum_{\mathbf{x}} \mathbb{P}(\mathbf{x})|Q(\mathbf{x})|)^{2/3}(\sum_{\mathbf{x}} \mathbb{P}(\mathbf{x})|Q(\mathbf{x})|^4)^{\frac{1}{3}} .\end{aligned}$$

Taking both sides of the above to the power of $3/2$ and rearranging, we obtain that,

$$\mathbb{E}[\,|Q(\mathbf{x})|\,] \geq \mathbb{E}[\,|Q(\mathbf{x})|^2\,]^{\frac{1}{2}} \left(\mathbb{E}[\,|Q(\mathbf{x})|^2\,]^{\frac{1}{2}}/\mathbb{E}[\,|Q(\mathbf{x})|^4\,]^{\frac{1}{4}}\right)^2 . \tag{8}$$

We now use Corollary (3.2) from [7] to get that $\mathbb{E}[|Q(\mathbf{x})|^2]^{\frac{1}{2}} \geq \sigma_{4,2}(\alpha)^d\mathbb{E}[|Q(\mathbf{x})|^4]^{\frac{1}{4}}$, where $\sigma_{4,2}(\alpha) = \sqrt{\frac{(1-\alpha)^{2/4}-\alpha^{2/4}}{(1-\alpha)\alpha^{2/4-1}-\alpha(1-\alpha)^{2/4-1}}}$. We conclude our proof by combining the above with Eq. 8 and noting that for $\alpha \in (.05, .5)$ we have $\sigma_{4,2}(\alpha)^2 \geq 0.2$. □

Lemma 5 *Let X be a random variable and $x' \in \mathbb{R}$ be a scalar and assume that there exists $a > 0$ such that for all $t \geq 0$ we have $\mathbb{P}[|X - x'| > t] \leq 2e^{-t^2/a^2}$. Then, $\mathbb{E}[|X - x'|] \leq 4\,a$.*

Proof For all $i = 0, 1, 2, \ldots$ denote $t_i = a\,i$. Since t_i is monotonically increasing we have that $\mathbb{E}[|X - x'|]$ is at most $\sum_{i=1}^{\infty} t_i\mathbb{P}[|X - x'| > t_{i-1}]$. Combining the above with the assumption in the lemma we get that $\mathbb{E}[|X - x'|] \leq 2\,a\,\sum_{i=1}^{\infty} i\exp^{-(i-1)^2}$. The proof now follows from the inequalities

$$\sum_{i=1}^{\infty} i\exp^{-(i-1)^2} \leq \sum_{i=1}^{5} i\exp^{-(i-1)^2} + \int_5^{\infty} xe^{-(x-1)^2}dx < 1.8 + 10^{-7} < 2 .$$

B Analysis of Example from Section 4.2

Optimal bounded ℓ_1-norm predictor: We first establish that $\min_{\|\mathbf{w}\|_1\leq 3}\mathbb{E}[|\,\langle\mathbf{w},\mathbf{x}\rangle - \mathsf{y}|] \leq l$. We do so by presenting a predictor with $\|\mathbf{w}\|_1 < 3$ and mean absolute-error at most l (this is *not* the optimal predictor, but certainly bound its mean absolute-error). Consider the predictor with $w_{i,1} = \frac{8}{3n}$ for each group i, and $w_{i,j} = 0$ for all $j > 1$. We have $\mathbb{E}[\langle\mathbf{w},\mathbf{x}\rangle\,|y] = \frac{8}{3n}\sum_{i=1}^{n}\frac{3}{4}(\frac{1}{3}+\frac{i-1}{3(n-1)})y = y$ and $\mathrm{Var}[\,\langle\mathbf{w},\mathbf{x}\rangle \mid \mathsf{y}] = \sum_i(\frac{8}{3n})^2(1-(\frac{3(1+(i-1))}{4\cdot 3})^2) \leq n(\frac{8}{3n})^2(1-(\frac{3\cdot 1}{4\cdot 3})^2) = \frac{20}{3n} < 7/n$. Using a calculation as in Eq. 6, we can now bound the mean absolute-error by $\sqrt{7/n} \leq l$.

Optimal predictor $\mathbf{w}^*$ with bounded ℓ_1-norm: Within each group, if $w^*_{i,j} > w^*_{i,j'}$ we can reduce the mean-absolute error by shifting weight from (i, j) to (i, j'). We can conclude that the weights within each group are equal. Furthermore, if $w^*_{i,1} \geq w^*_{n,1}$ and $w^*_{i,1} > 0$, for $i < n$, we can reduce the error by shifting weight from group i to the last group. Therefore, for any non-zero bound on the ℓ_1-norm, the last group would have strictly more weight than other groups.

Predictor using only last group: Consider a predictor $\mathbf{w}$ that uses only the last group, i.e. with $w_{i,j} = 0$ for $i < n$. We have $\mathbb{E}[\langle\mathbf{w},\mathbf{x}\rangle\,|y] = \sum_j w_{n,j}\frac{3}{4}\frac{2}{3}y = \frac{1}{2}\rho y$, where $\rho = \sum_j w_{n,j}$, and $\mathrm{Var}[\,\langle\mathbf{w},\mathbf{x}\rangle \mid \mathsf{y}] = \sum_i w^2_{n,i}\mathrm{Var}[\mathsf{x}_{n,i} \mid y] + \sum_{i\neq j} w_{n,i}w_{n,j}\mathrm{Cov}[\mathsf{x}_{n,i},\mathsf{x}_{n,j} \mid y] \geq (\sum_i w_{n,i})^2\mathrm{Cov}[\mathsf{x}_{n,1},\mathsf{x}_{n,2} \mid y] = \frac{5}{16}\rho^2$. In order to apply Lemma 4 we introduce i.i.d., and independent of $\mathbf{z}$, latent Bernoulli variables $\mathsf{v}_{n,j}$, with $\mathbb{P}(\mathsf{v}_{n,j} = 1) = \frac{7}{8}$, and write $\mathsf{x}_{n,j} = \mathsf{z}_n\mathsf{v}_{n,j}$.

The mean absolute-error is now a second degree polynomial in the independent (conditioned on y) variables $\mathbf{v}, \mathbf{z}$, and we can apply Lemma 4 and use arguments as in the Proof of 2 to get $\mathbb{E}[|\langle \mathbf{w}, \mathbf{x} \rangle - \mathsf{y}|] \geq 0.2^2 \sqrt{\frac{5}{16}\rho^2 + (\frac{1}{2}\rho - 1)^2} \geq 0.2^2 \sqrt{\frac{45}{81}} > 0.02$.