TRADING ACCURACY FOR SPARSITY IN OPTIMIZATION PROBLEMS WITH SPARSITY CONSTRAINTS

SHAI SHALEV-SHWARTZ*, NATHAN SREBRO†, AND TONG ZHANG‡

Abstract. We study the problem of minimizing the expected loss of a linear predictor while constraining its sparsity, i.e., bounding the number of features used by the predictor. While the resulting optimization problem is generally NP-hard, several approximation algorithms are considered. We analyze the performance of these algorithms, focusing on the characterization of the trade-off between accuracy and sparsity of the learned predictor in different scenarios.

1. Introduction. In statistical and machine learning applications, although 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.

These applications lead to optimization problems with sparsity constraints. Focusing on linear prediction, it is generally NP-hard to find the best predictor subject to a sparsity constraint, i.e. a bound on the number of features used (19; 7). In this paper we show that by compromising on prediction accuracy, one can compute sparse predictors efficiently. Our main goal is to understand the precise trade-off between accuracy and sparsity, and how this trade-off depends on properties of the underlying optimization problem.

We now formally define our problem setting. A linear predictor is a mapping $\mathbf{x} \mapsto \phi(\langle \mathbf{w}, \mathbf{x} \rangle)$ where $\mathbf{x} \in \mathcal{X} \stackrel{\text{def}}{=} [-1, +1]^d$ is a d-dimensional vector of features, $\mathbf{w} \in \mathbb{R}^d$ is the linear predictor, $\langle \mathbf{w}, \mathbf{x} \rangle$ is the inner-product operation, and $\phi : \mathbb{R} \to \mathcal{Y}$ is a scalar function that maps the scalar $\langle \mathbf{w}, \mathbf{x} \rangle$ to the desired output space $\mathcal{Y}$. For example, in binary classification problems we have $\mathcal{Y} = \{-1, +1\}$ and a linear classification rule is $\mathbf{x} \mapsto \text{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle)$. In regression problems, $Y = \mathbb{R}$, ϕ is the identity function, and the linear regression rule is $\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$.

The loss of a linear predictor $\mathbf{w}$ on an example $(\mathbf{x}, y)$ is assessed by a loss function $L(\langle \mathbf{w}, \mathbf{x} \rangle, y)$. Note that ϕ does not appear in the above expression. This is convenient since in some situations the loss also depends on the pre-image of ϕ. For example, the hinge-loss that is used in Support Vector Machines (29) is defined as $L(\langle \mathbf{w}, \mathbf{x} \rangle, y) = \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x} \rangle\}$. Other notable examples of loss functions are the squared loss, $L(\langle \mathbf{w}, \mathbf{x} \rangle, y) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$, the absolute loss, $L(\langle \mathbf{w}, \mathbf{x} \rangle, y) = |\langle \mathbf{w}, \mathbf{x} \rangle - y)|$, and the logistic loss, $L(\langle \mathbf{w}, \mathbf{x} \rangle, y) = \log(1 + \exp(-y\langle \mathbf{w}, \mathbf{x} \rangle))$. Throughout this paper, we always assume:

ASSUMPTION 1.1. *$L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is convex with respect to its first argument.*

Given a joint distribution over $\mathcal{X} \times \mathcal{Y}$, the risk of a linear predictor $\mathbf{w}$ is its expected

*School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel shais@cs.huji.ac.il.

†Toyota Technological Institute, Chicago, USA, nati@tti-c.org.

‡Statistics Department Rutgers University, NJ, USA, tzhang@stat.rutgers.edu. This author is partially supported by the following grants: NSF DMS-1007527, NSA-081024, and AFOSR-10097389.

loss:

$$R(\mathbf{w}) \;=\; \mathbb{E}_{(\mathbf{x},y)}[L(\langle \mathbf{w}, \mathbf{x}\rangle, y)] \;. \tag{1.1}$$

Often, we do not know the distribution over examples, but instead approximate it using the uniform probability over a finite training set. In that case, the empirical risk is $\frac{1}{m}\sum_{i=1}^{m} L(\langle \mathbf{w}, \mathbf{x}_i\rangle, y_i)$. Since this expression is a special case of the objective given in Equation (1.1), we stick to the more general definition given in Equation (1.1). In some situations, one can add regularization to the empirical risk. We discuss regularized risk in later sections.

The *sparsity* of a linear predictor $\mathbf{w}$ is defined to be the number of non-zero elements of $\mathbf{w}$ and denoted using the ℓ_0 notation:

$$\|\mathbf{w}\|_0 = |\{i : w_i \neq 0\}| \;. \tag{1.2}$$

In this paper, we study the problem of computing sparse predictors. That is, we are interested in linear predictors that on one hand achieve low risk while on the other hand have low ℓ_0 norm. Sometime these two goals are contradictory. Thus, we would like to understand the trade-off between $R(\mathbf{w})$ and $\|\mathbf{w}\|_0$. Ultimately, to balance the trade-off, one aims at (approximately) solving the following sparsity constrained optimization problem:

$$\min_{\mathbf{w}:\|\mathbf{w}\|_0 \leq B_0} R(\mathbf{w}) \;. \tag{1.3}$$

That is, find the predictor with minimal risk among all predictors with ℓ_0 norm bounded by some sparsity parameter B_0. Regrettably, the constraint $\|\mathbf{w}\|_0 \leq B_0$ is non-convex, and solving the optimization problem in Equation (1.3) is NP-hard (19; 7).

To overcome the hardness result, two main approaches have been proposed in the literature. In the first approach, we replace the non-convex constraint $\|\mathbf{w}\|_0 \leq B_0$ with the convex constraint $\|\mathbf{w}\|_1 \leq B_1$. The problem now becomes a convex optimization problem that can be solved efficiently. It was observed that the ℓ_1 norm sometimes encourages sparse solutions. But, it can be shown that the solution of an ℓ_1 constrained risk minimization is not always sparse. Moreover, it is important to quantify the sparsity one can obtain using the ℓ_1 relaxation.

The second approach for overcoming the hardness of solving Equation (1.3) is called forward greedy selection (a.k.a. Boosting). In this approach, we start with the all-zeros predictor, and at each step we change a single element of the predictor in a greedy manner, so as to maximize the decrease in risk due to this change. Here, early stopping guarantees a bound on the sparsity level of the output predictor. But, the price of early stopping is a sub-optimal accuracy of the resulting predictor (i.e. the risk, $R(\mathbf{w})$, can be high).

In this paper we study trade-offs between accuracy and sparsity for ℓ_0 or ℓ_1 bounded predictors. We provide results to answer the following questions: Given an excess risk parameter ϵ, for what sparsity level B, can we *efficiently* find a predictor with sparsity level $\|\mathbf{w}\|_0 \leq B$ and risk bounded by $R(\bar{\mathbf{w}}) + \epsilon$, where $\bar{\mathbf{w}}$ is an unknown reference predictor? Moreover, how does B depend on ϵ, and on properties of the loss function L, the distribution over examples, and the reference vector $\bar{\mathbf{w}}$?

1.1. Additional notation and definitions. The set of integers $\{1, \ldots, d\}$ is denoted by $[d]$. For a vector $\mathbf{w}$, the support of $\mathbf{w}$ is defined to be: $\text{supp}(\mathbf{w}) = \{i \in$

$[d] : w_i \neq 0\}$. For $i \in [d]$, the vector $\mathbf{e}^i$ is the all zeros vector except 1 in the ith element.

The following definitions characterize two types of loss functions.

DEFINITION 1.1 (Lipschitz loss). *A loss function $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is ρ-Lipschitz continuous if*

$$\forall y \in \mathcal{Y},\ \forall a, b \quad |L(a, y) - L(b, y)| \leq \rho |a - b| \ .$$

Examples of 1-Lipschitz loss functions are the hinge loss, $L(a, y) = \max\{0, 1 - ya\}$ and the absolute loss, $L(a, y) = |a - y|$. See Section 3 for details.

DEFINITION 1.2 (Smooth Loss). *A loss function $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is β-smooth if*

$$\forall y, \forall a, b \quad L(a, y) - L(b, y) \leq L'(b, y)\,(a - b) + \frac{\beta\,(a - b)^2}{2} \quad ,$$

where $L'(b, y)$ is the derivative of L with respect to its first argument at (b, y).

In Lemma B.1 we also show how this translates into a smoothness property of the risk function, $R(\mathbf{w})$. Examples of smooth losses are the logistic loss and the quadratic loss. See Section 3 for details.

Finally, the following definition characterizes properties of the risk function.

DEFINITION 1.3 (Strong convexity). *$R(\mathbf{w})$ is said to be λ-strongly convex if*

$$\forall \mathbf{w}, \mathbf{u}, \quad R(\mathbf{w}) - R(\mathbf{u}) - \langle \nabla R(\mathbf{u}), \mathbf{w} - \mathbf{u} \rangle \geq \frac{\lambda}{2} \|\mathbf{w} - \mathbf{u}\|_2^2 \quad .$$

Similarly, $R(\mathbf{w})$ is λ-strongly convex on a set $F \subset [d]$ if the above inequality holds for all $\mathbf{w}, \mathbf{u}$ such that $\mathrm{supp}(\mathbf{w}) \subseteq F$ and $\mathrm{supp}(\mathbf{u}) \subseteq F$. Finally, $R(\mathbf{w})$ is (k, λ)-sparsely-strongly convex if for any $F \subset [d]$ such that $|F| \leq k$, $R(\mathbf{w})$ is λ-strongly convex on F.

1.2. Outline. The paper is organized as follows. In Section 2 we describe our main results, showing the tradeoff between sparsity and accuracy in different scenarios. Several examples are mentioned in Section 3. Next, in Section 4 we formally show that some of the relation between accuracy and sparsity outlined in Section 2 are tight. In Section 5 we put our work in context and review related work. In particular, we show that the algorithms we present in Section 2 for studying the tradeoff between sparsity and accuracy are variants of previously proposed algorithms. Our main contribution is the systematic analysis of the tradeoff between sparsity and accuracy. Despite the fact that some of our results can be derived from previous work, for the sake of completeness and clarity, we provide complete proofs of all our results in Section A.

2. Main Results. We now state our main findings. We present four methods for computing sparse predictors. In the first method, we first solve the ℓ_1 relaxed problem and then use randomization for sparsifying the resulting predictor. In the second method, we take a more direct approach and describe a forward greedy selection algorithm for incrementally solving the ℓ_1 relaxed problem. For this approach, we show how early stopping provides a trade-off between sparsity and accuracy. Finally, in the last two methods we do not use the ℓ_1 constraint at all, but only rely on early stopping of another greedy method. We show that these methods are guaranteed to be comparable to the other methods and sometimes they significantly outperform the other methods. They also has the advantage of not relying on any parameters.

2.1. Randomized sparsification of low ℓ_1 predictors. In the ℓ_1 relaxation approach, we first solve the problem

$$\min_{\mathbf{w}:\|\mathbf{w}\|_1 \leq B_1} R(\mathbf{w}) \ . \tag{2.1}$$

Let $\mathbf{w}^\star$ be an optimal solution of Equation (2.1). Although $\mathbf{w}^\star$ may be sparse in some situations, in general we have no guarantees on $\|\mathbf{w}^\star\|_0$. Our goal is to find a sparse approximation of $\mathbf{w}^\star$, while not paying too much in the risk value. A simple way to do this is to use the following randomized sparsification procedure, which was originally proposed by Maurey (22). See also Section 5 for additional references.

Algorithm 1 Randomized Sparsification

INPUT: vector $\mathbf{w}^\star \in \mathbb{R}^d$
let $\mathbf{w}^{(0)} = 0$
for $k = 1, 2, \ldots$
  sample $r_k \in [d]$ according to
    the distribution $\Pr[r_k = j] = |w^\star_j| / \|\mathbf{w}^\star\|_1$
  let $\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} + \text{sign}(w^\star_{r_k})\, \mathbf{e}^{r_k}$
end
OUTPUT: $\frac{\|\mathbf{w}^\star\|_1}{k} \mathbf{w}^{(k)}$

Intuitively, we view the prediction $\langle \mathbf{w}, \mathbf{x} \rangle$ as the expected value of the elements in $\mathbf{x}$ according to the distribution vector $\mathbf{w}/\|\mathbf{w}\|_1$. The randomized sparsification procedure approximate this expected value by randomly selecting elements from $[d]$ according to the probability measure $\mathbf{w}/\|\mathbf{w}\|_1$.

Clearly, if we run the sparsification procedure for k iterations we have $\|\mathbf{w}^{(k)}\|_0 \leq k$. The following theorem shows how the excess risk of $\mathbf{w}^{(k)}$ depends on k and on $\|\mathbf{w}^\star\|_1$. The bounds are not new and special cases of this theorem have been derived before in (14; 2; 16; 24; 27; 6).

THEOREM 2.1. *Let $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a loss function and let $R(\mathbf{w})$ be as defined in Equation (1.1), where the expectation is w.r.t. an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\mathbf{w}^\star$ be the input of the randomized sparsification procedure (Algorithm 1) and let $\mathbf{w}$ be its output after performing k iterations. Then, for any $\epsilon > 0$, with probability of at least $1/2$ over the choice of $r_1, \ldots, r_k$ we have $R(\mathbf{w}) - R(\mathbf{w}^\star) \leq \epsilon$ provided that:*

$$k \ \geq \ \begin{cases} 2\,\frac{\rho^2 \|\mathbf{w}^\star\|_1^2}{\epsilon^2} & \text{if } L \text{ is } \rho \text{ Lipschitz} \\ \frac{\beta \|\mathbf{w}^\star\|_1^2}{\epsilon} & \text{if } L \text{ is } \beta \text{ smooth} \end{cases}$$

The above theorem implies that on average, if we repeat the randomized procedure twice and choose the $\mathbf{w}$ with minimal risk then we obtain $R(\mathbf{w}) - R(\mathbf{w}^\star) \leq \epsilon$. Furthermore, for any $\delta \in (0, 1)$, if we repeat the randomized procedure $\lceil \log(1/\delta) \rceil$ times and choose the $\mathbf{w}$ with minimal risk, then the probability that $R(\mathbf{w}) - R(\mathbf{w}^\star) > \epsilon$ is at most δ.

Let $\bar{\mathbf{w}} \in \mathbb{R}^d$ be an arbitrary (unknown) predictor. The guarantees given in Theorem 2.1 tell us that if $\|\bar{\mathbf{w}}\|_1 = B_1$ then we can find a predictor with $R(\mathbf{w}) - R(\bar{\mathbf{w}}) \leq \epsilon$ provided that k is sufficiently large, and the lower bound on k depends on the ℓ_1 norm of $\bar{\mathbf{w}}$. We next show that by assuming more about the risk function, we can have a result that involves the ℓ_0 norm of the reference vector. In particular, we will assume

that the risk is strongly convex on the support of $\bar{\mathbf{w}}$. The importance of strongly convex risk in this context stems from the following lemma in which we show that if the risk is strongly convex then $\|\bar{\mathbf{w}}\|_1^2$ can be bounded using the ℓ_0 norm of $\bar{\mathbf{w}}$ and the strong convexity parameter.

LEMMA 2.2. *Let $F \subset [d]$ and assume that $R(\mathbf{w})$ is λ-strongly convex on F. Let*

$$\bar{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}:\mathrm{supp}(\mathbf{w})=F} R(\mathbf{w}) \ .$$

Then,

$$\|\bar{\mathbf{w}}\|_1 \leq \sqrt{\frac{2\,\|\bar{\mathbf{w}}\|_0\,(R(\mathbf{0}) - R(\bar{\mathbf{w}}))}{\lambda}} \ .$$

Combining the above lemma with Theorem 2.1 we immediately get:

COROLLARY 2.3. *Let $F \subset [d]$ and assume that $R(\mathbf{w})$ is λ-strongly convex on F. Let*

$$\bar{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}:\mathrm{supp}(\mathbf{w})=F} R(\mathbf{w}) \ ,$$

let $\mathbf{w}^\star$ be a minimizer of Equation (2.1) with $B_1 = \sqrt{\frac{2\,\|\bar{\mathbf{w}}\|_0\,(R(\mathbf{0})-R(\bar{\mathbf{w}}))}{\lambda}}$, and let $\mathbf{w}$ be the output of the randomized sparsification procedure (Algorithm 1). Then, for any $\epsilon > 0$, with probability of at least 0.5 over the choice of $r_1, \ldots, r_k$ we have $R(\mathbf{w}) - R(\bar{\mathbf{w}}) \leq \epsilon$ provided that the following holds:

$$k \ \geq \ \begin{cases} \|\bar{\mathbf{w}}\|_0 \, \frac{4\,\rho^2\,(R(\mathbf{0})-R(\bar{\mathbf{w}}))}{\lambda\,\epsilon^2} & \text{if } L \text{ is } \rho \text{ Lipschitz} \\ \|\bar{\mathbf{w}}\|_0 \, \frac{2\,\beta\,(R(\mathbf{0})-R(\bar{\mathbf{w}}))}{\lambda\,\epsilon} & \text{if } L \text{ is } \beta \text{ smooth} \end{cases}$$

In Section 3 we demonstrate cases in which the conditions of Corollary 2.3 holds. Note that we have two means to control the trade-off between sparsity and accuracy. First, using the parameter ϵ. Second, using the reference vector $\bar{\mathbf{w}}$, since by choosing $\bar{\mathbf{w}}$ for which the risk is strongly convex on $\mathrm{supp}(\bar{\mathbf{w}})$ we obtain better sparsity guarantee, but the price we pay is that this restriction might increase the risk of $\bar{\mathbf{w}}$. For more details see Section 3.

2.2. Forward Greedy Selection. The approach described in the previous subsection involves two steps. First, we solve the ℓ_1 relaxed problem given in Equation (2.1) and only then we apply the randomized sparsification procedure. In this section we describe a more direct approach in which we solve Equation (2.1) using an iterative algorithm that alters a single element of $\mathbf{w}$ at each iteration. We derive upper bounds on the number of iterations required to achieve an ϵ accurate solution, which immediately translates to bounds on the sparsity of the approximated solution. Variants of the algorithm below and its analysis were proposed before by several authors (10; 32; 6). The version we have here includes closed form definition of the step size and a stopping criterion that depends on the desired accuracy ϵ.

Algorithm 2 Forward Greedy Selection

PARAMETERS: positive scalars B_1, ϵ
let $\mathbf{w}^{(0)} = 0$
for $k = 0, 1, 2, \ldots$
  let $\boldsymbol{\theta}^{(k)} = \nabla R(\mathbf{w}^{(k)})$
  let $r_k = \operatorname{argmax}_j |\theta_j^{(k)}|$
  let $\eta_k = \min\left\{1, \frac{(\langle \boldsymbol{\theta}^{(k)}, \mathbf{w}^{(k)} \rangle + B_1 \|\boldsymbol{\theta}^{(k)}\|_\infty)}{4\, B_1^2\, \beta}\right\}$
  let $\mathbf{w}^{(k+1)} = (1 - \eta_k)\mathbf{w}^{(k)} + \eta_k \operatorname{sgn}(-\theta_{r_k}^{(k)})\, B_1\, \mathbf{e}^{r_k}$
STOPPING CONDITION: $\langle \boldsymbol{\theta}^{(k)}, \mathbf{w}^{(k)} \rangle + B_1 \|\boldsymbol{\theta}^{(k)}\|_\infty \leq \epsilon$

The algorithm initializes the predictor vector to be the zero vector, $\mathbf{w}^{(1)} = \mathbf{0}$. On iteration k, we first choose a feature by calculating the gradient of R at $\mathbf{w}^{(k)}$ (denoted $\boldsymbol{\theta}^{(k)}$) and finding its largest element in absolute value. Then, we calculate a step size η_k and update the predictor to be a convex combination of the previous predictor and the singleton $B_1\, \mathbf{e}^{r_k}$ (with appropriate sign). The step size and the stopping criterion are based on our analysis. Note that the update form ensures us that for all k, $\|\mathbf{w}^{(k)}\|_1 \leq B_1$ and $\|\mathbf{w}^{(k)}\|_0 \leq k$.

The following theorem upper bounds the number of iterations required by the Forward Greedy Selection algorithm. The theorem holds for the case of smooth loss functions. As mentioned previously, variants of this theorem have been given in (10; 32; 6).

THEOREM 2.4. *Let $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a convex β-smooth loss function and let $R(\mathbf{w})$ be as defined in Equation (1.1), where the expectation is w.r.t. an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose that the Forward Greedy Selection procedure (Algorithm 2) is run with parameters B_1, ϵ and let $\mathbf{w}^\star$ be a minimizer of Equation (2.1). Then, the algorithm terminates after at most*

$$k \;\leq\; \left\lceil \frac{8\, \beta\, B_1^2}{\epsilon} \right\rceil$$

iterations, and at termination, $R(\mathbf{w}^{(k)}) - R(\mathbf{w}^\star) \leq \epsilon$.

Since a bound on the number of iterations of the Forward Greedy Selection algorithm translates into a bound on the sparsity of the solution, we see that the guarantee we obtain from Theorem 2.4 is similar to the guarantee we obtain from Theorem 2.1 for the randomized sparsification. The advantages of the direct approach we take here is its simplicity – we do not need to solve Equation (2.1) in advance and we do not need to rely on randomization.

Next, we turn to derivation of the sparsification result for the case that L is ρ-Lipschitz but is not β-smooth. To do so, we approximate L by a β-smooth function. This can always be done, as the following lemma indicates.

LEMMA 2.5. *Let L be a proper, convex, ρ-Lipschitz loss function and let $\tilde{L}$ be defined as follows*

$$\forall y \in \mathcal{Y},\; \tilde{L}(a, y) \;=\; \inf_v \left[\frac{\beta}{2} v^2 + L(a - v, y) \right] . \tag{2.2}$$

Then, $\tilde{L}$ is β-smooth and

$$\forall y \in \mathcal{Y}, a \in \mathbb{R}, \quad 0 \leq L(a, y) - \tilde{L}(a, y) \leq \frac{\rho^2}{2\, \beta} \quad .$$

Let $\tilde{R}(\mathbf{w}) = \mathbb{E}[\tilde{L}(\langle \mathbf{w}, \mathbf{x} \rangle, y)]$. Clearly, for all $\mathbf{w}$ we have $0 \leq R(\mathbf{w}) - \tilde{R}(\mathbf{w}) \leq \frac{\rho^2}{2\beta}$. As a direct corollary we obtain:

COROLLARY 2.6. *Let $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a ρ-Lipschitz convex loss function and let $R(\mathbf{w})$ be as defined in Equation (1.1), where the expectation is w.r.t. an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose that the Forward Greedy Selection procedure (Algorithm 2) is run with parameters B_1, ϵ on the function $\tilde{R}(\mathbf{w}) = \mathbb{E}[\tilde{L}(\langle \mathbf{w}, \mathbf{x} \rangle, y)]$, where $\tilde{L}$ is as defined in Equation (2.2) and $\beta = \frac{\rho^2}{\epsilon}$. Then, the algorithm stops after at most*

$$k \;\leq\; \left\lceil \frac{8\,\rho^2\,B_1^2}{\epsilon^2} \right\rceil$$

iterations, and when it stops we have $R(\mathbf{w}^{(k)}) - R(\mathbf{w}^\star) \leq \epsilon$, where $\mathbf{w}^\star$ is a minimizer of Equation (2.1).

The above corollary gives a similar guarantee to the one given in Theorem 2.1 for the case of Lipschitz loss functions.

Finally, if the risk function is strongly convex on the support of a vector $\bar{\mathbf{w}}$, we can obtain the same guarantee as in Corollary 2.3 for the Forward Greedy Selection algorithm by combining Theorem 2.4 and Corollary 2.6 with the bound on the ℓ_1 norm of $\bar{\mathbf{w}}$ given in Lemma 2.2.

To summarize this subsection, we have shown that the Forward Greedy Selection procedure provides the same guarantees as the method which first solves the ℓ_1 relaxed problem and then uses randomized sparsification. The Forward Greedy Selection procedure is a deterministic, more direct, simple, and efficient approach. In the next subsection we provide an even better method.

2.3. Fully Corrective Greedy Selection. The Forward Greedy Selection method described in the previous subsection is a nice and simple approach. However, intuitively, this method is wasteful since at each iteration, we may increase the support of the solution, although it is possible that we can reduce the risk by only modifying the weights of the current support. It makes sense to first fully adjust the weights of the current features so as to minimize the risk, and only then add a fresh feature to the support of the solution. In this subsection we present our last method, which exactly do this. In addition, the new method do not enforce the constraint $\|\mathbf{w}\|_1 \leq B_1$ at all. This stands in contrast to the two methods described previously in which we are required to tune the parameter B_1 in advance. Nevertheless, as we will show below, the new method achieves the same guarantees as the previous methods and sometime it even achieves improved guarantees. At the end of the subsection, we present additional post-processing procedure which does not modify the sparsity of the solution but may improve its accuracy.

Algorithm 3 Fully Corrective Forward Greedy Selection

let $\mathbf{w}^{(0)} = 0$
let $F^{(0)} = \emptyset$
for $k = 1, 2, \ldots$
  let $r_k = \operatorname{argmin}_j \min_\alpha R(\mathbf{w}^{(k-1)} + \alpha \mathbf{e}^j)$
  let $F^{(k)} = F^{(k-1)} \cup \{r_k\}$
  let $\mathbf{w}^{(k)} = \operatorname{argmin}_{\mathbf{w}} R(\mathbf{w})$ s.t. $\operatorname{supp}(\mathbf{w}) \subseteq F^{(k)}$
end

The Fully Corrective algorithm is similar to the non corrective algorithm described in the previous subsection with two main differences. First, in Algorithm 3 we adjust the weights so as to minimize the risk over the features aggregated so far. This is what we mean by *fully corrective*. Second, we now do not enforce the constraint $\|\mathbf{w}\|_1 \leq B_1$.

Although in Algorithm 3 we choose r_k to be the feature which leads to the largest decrease of the risk, from the proof, we can see that identical results hold by choosing

$$r_k = \operatorname*{argmax}_j |\nabla R(\mathbf{w}^{(k)})_j|$$

as in Algorithm 2. Moreover, if $R(\mathbf{w})$ can be represented as $R(\mathbf{w}) = Q(X\mathbf{w})$, where each row of the matrix X is one example, and let X^j be the j-th column of X, with normalization $\langle X^j, X^j \rangle = 1$, then we may also choose r_k to optimize the quadratic approximation function

$$r_k = \operatorname*{argmin}_j \min_\alpha \left\| \alpha X^j - \nabla Q(X\mathbf{w}^{(k-1)}) \right\|_2^2,$$

and again, identical results hold. For prediction problems, this formulation leads to the Fully Corrective version of the functional gradient boosting method (12), where this quadratic approximation is equivalent to a regression problem.

We now turn to the analysis of the Fully Corrective algorithm. Our first theorem provides a similar guarantee to the one given in Theorem 2.4. However, as mentioned before, the Fully Corrective algorithm is parameter free, and therefore we obtain a guarantee which holds simultaneously for all values of B_1.

THEOREM 2.7. *Let $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a convex β-smooth loss function and let $R(\mathbf{w})$ be as defined in Equation (1.1), where the expectation is w.r.t. an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose that the Fully Corrective procedure (Algorithm 3) is run for k iterations. Then, for any scalar $\epsilon > 0$ and vector $\bar{\mathbf{w}}$ such that*

$$k \;\geq\; \frac{2\,\beta\,\|\bar{\mathbf{w}}\|_1^2}{\epsilon}$$

we have $R(\mathbf{w}^{(k)}) - R(\bar{\mathbf{w}}) \leq \epsilon$.

Naturally, if our loss function is Lipschitz but is not smooth, we can run the Fully Corrective algorithm on the modified loss $\tilde{L}$ (see Lemma 2.5) and obtain a guarantee similar to the one in Corollary 2.6. Similarly, if the risk function is strongly convex on the support of $\bar{\mathbf{w}}$, we can use Lemma 2.2 to obtain the same guarantee as in Corollary 2.3. Therefore, we have shown that the Fully Corrective method provides the same guarantees as the previous approaches, with the important advantage that B_1 appears only in the analysis but does not affect the algorithm.

Finally, we show that with a more restricted assumption on the risk function, we can obtain an exponentially better dependence on $\frac{1}{\epsilon}$ for the Fully Corrective algorithm.

THEOREM 2.8. *Let $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a convex β-smooth loss function and let $R(\mathbf{w})$ be as defined in Equation (1.1), where the expectation is w.r.t. an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose that the Fully Corrective procedure (Algorithm 3) is run for k iterations. Let $\lambda > 0$ be a scalar and assume that R is $(k + \|\bar{\mathbf{w}}\|_0, \lambda)$-sparsely-strongly convex. Then, for any $\epsilon > 0$, and $\bar{\mathbf{w}} \in \mathbb{R}^d$ such that*

$$k \;\geq\; \|\bar{\mathbf{w}}\|_0 \,\frac{\beta}{\lambda}\, \log\left(\tfrac{R(\mathbf{0}) - R(\bar{\mathbf{w}})}{\epsilon}\right) \quad ,$$

we have $R(\mathbf{w}^{(k)}) \leq R(\bar{\mathbf{w}}) + \epsilon$.

REMARK 2.1. *It is possible to show that the result of Theorem 2.8 still holds if we use an ℓ_2 regularized risk, that is, define $R(\mathbf{w}) = \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$. Note that in this case, the optimal solution will in general be dense, since the ℓ_2 regularization tends to spread the weights of the solution over many features. However, since Theorem 2.8 holds for* any *reference vector $\bar{\mathbf{w}}$, and not only for the minimizer of the risk, it suffices that there will be some sparse vector $\bar{\mathbf{w}}$ that achieves a low risk. In this case, Theorem 2.8 guarantees that we will find $\mathbf{w}^{(k)}$ whose risk is only slightly higher than that of $\bar{\mathbf{w}}$ and whose sparsity is only slightly worse than $\bar{\mathbf{w}}$.*

Adding Replacement Steps as Post-Processing. We can always try to improve the solution without changing its sparsity level. The following procedure suggests one way how to do this. The basic idea is simple. We first perform one Fully Corrective forward selection step, and second we remove the feature that has the smallest weight. We accept such a replacement operation only if it leads to a smaller value of $R(\mathbf{w})$. The resulting procedure is summarized in Algorithm 4.

Algorithm 4 Post Processing Replacement Steps

INPUT: $F^{(0)} \subset [d]$
for $t = 0, 1, \ldots$
  let $\mathbf{w}^{(t)} = \operatorname{argmin}_{\mathbf{w}} R(\mathbf{w})$ s.t. $\operatorname{supp}(\mathbf{w}) \subseteq F^{(t)}$
  let $r_{t+1} = \operatorname{argmin}_j \min_\alpha R(\mathbf{w}^{(t)} + \alpha \mathbf{e}^j)$
  let $F' = F^{(t)} \cup \{r_{t+1}\}$
  let $\mathbf{w}' = \operatorname{argmin}_{\mathbf{w}} R(\mathbf{w})$ s.t. $\operatorname{supp}(\mathbf{w}) \subseteq F'$
  let $q = \operatorname{argmin}_{j \in F'} |w'_j|$
  let $\delta_t = R(\mathbf{w}^{(t)}) - R(\mathbf{w}' - w'_q \mathbf{e}^q)$
  if $(\delta_t \leq 0)$ **break**
  let $F^{(t+1)} = F' - \{q\}$
end

We may also use a slightly simpler replacement procedure that skips the optimization step in F'. That is, we simultaneously take

$$r_{t+1} = \operatorname*{argmax}_j |\nabla R(\mathbf{w}^{(t)})_j|, \quad q = \operatorname*{argmin}_{j \in F^{(k)}} |\mathbf{w}_j^{(t)}|,$$

and let $F^{(t+1)} = (F^{(t)} \cup \{r_{t+1}\}) - \{q\}$. Similar results hold for this alternative.

Clearly, Algorithm 4 can only improve the objective. Thus, for any t we have $R(\mathbf{w}^{(t)}) \leq R(\mathbf{w}^{(0)})$. The following theorem states that we can have an actual decrease of R by running Algorithm 4 as a post processing to Algorithm 3.

THEOREM 2.9. *Let $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a convex β-smooth loss function and let $R(\mathbf{w})$ be as defined in Equation (1.1), where the expectation is w.r.t. an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\lambda > 0$ be a scalar, k be an integer, and $\bar{\mathbf{w}} \in \mathbb{R}^d$ be a vector, such that*

$$k + 1 \;\geq\; \|\bar{\mathbf{w}}\|_0 (1 + 4\beta^2/\lambda^2) \quad ,$$

and assume that R is $(k + 1 + \|\bar{\mathbf{w}}\|_0, \lambda)$-sparsely-strongly convex. Additionally, let T be an integer such that

$$T \geq \frac{\lambda(k + 1 - \|\bar{\mathbf{w}}\|_0)}{2\,\beta} \log\left(\tfrac{R(\mathbf{0}) - R(\bar{\mathbf{w}})}{\epsilon}\right) \quad .$$

Then, if the Fully Corrective procedure (Algorithm 3) is run for k iterations and its last predictor is provided as input for the post-processing Replacement procedure (Algorithm 4), which is then run for T iterations, then when the procedure terminates at time t (which may be smaller than T), we have $R(\mathbf{w}^{(t)}) - R(\bar{\mathbf{w}}) \leq \epsilon$.

The above theorem tells us that under the strong convexity condition, one may approximate $R(\bar{\mathbf{w}})$ to arbitrary precision using a number of features which is at most a constant $(1 + 4\beta^2/\lambda^2)$ approximation factor. Comparing this sparsity guarantee to the guarantee given in Theorem 2.8 we note that the sparsity level in Theorem 2.9 does not depend on $\log(1/\epsilon)$. Only the runtime depends on the desired accuracy level ϵ. In particular, if k is close to its lower bound, then the required number of iterations of Algorithm 4 becomes $O\left(\|\bar{\mathbf{w}}\|_0 \frac{\beta}{\lambda} \log(\frac{R(\mathbf{0})-R(\bar{\mathbf{w}})}{\epsilon})\right)$, which matches the bound on the number of iterations of Algorithm 3 given in Theorem 2.8. However, since Algorithm 4 does not increase the sparsity of the solution, decreasing ϵ solely translates to an increased runtime while not affecting the sparsity of the solution. On the flip side, the dependence of the sparsity of the solution on β/λ is linear in Theorem 2.8 and quadratic in Theorem 2.9.

It is worth pointing out that the result in Theorem 2.9 is stronger than results in the compressed sensing literature, which consider the least squares loss, and the bounds are of the flavor $R(\bar{\mathbf{w}}^{(t)}) \leq CR(\bar{\mathbf{w}})$ with some constant $C > 1$. For such a bound to be useful, we have to assume that $R(\bar{\mathbf{w}})$ is close to zero. This assumption is not needed in Theorem 2.9. However if we do assume that $R(\bar{\mathbf{w}})$ is close to the global minimum, then it is not difficult to see that $\bar{\mathbf{w}}^{(t)}$ is close to $\bar{\mathbf{w}}$ from the sparse strong convexity assumption. This implies a recovery result similar to those in compressed sensing. Therefore from the numerical optimization point of view, our analysis is more general than compressed sensing, and the latter may be regarded as a specialized consequence of our result.

Note that a more sophisticated combination of forward and backward updates is done by the FoBa algorithm of (33). The more aggressive backward steps in FoBa can lead to further improvement, in the sense that one may solve the sparse optimization problem exactly (that is, $\mathbf{w}^{(k)}$ contains only $k = \|\bar{\mathbf{w}}\|_0$ features). However, this requires additional assumptions. Most notably, it requires that $\bar{\mathbf{w}}$ will be the unique minimizer of $R(\mathbf{w})$. In contrast, in our case $\bar{\mathbf{w}}$ can be an arbitrary competing vector, a fact that gives us an additional control on the trade-off between sparsity and accuracy. See the discussion in Section 5 for more details.

We note that in practice it should be beneficiary to include some replacement steps during the entire run of the algorithm and not only as post processing steps. However, the analysis of such an algorithm is more complex because it depends on how to integrate the forward steps in Algorithm 3 and replacement steps in Algorithm 4. This paper only consider the simple situation that Algorithm 4 is run as a post-processing procedure, for which the theoretical result can be more easily stated and interpreted.

3. Examples. In this section we provide concrete examples that exemplify the usefulness of the bounds stated in the previous section.

We first list some loss functions.

Squared loss:. $L(a, y) = \frac{1}{2}(a - y)^2$. The domain $\mathcal{Y}$ is usually taken to be a bounded subset of $\mathbb{R}$. The second derivative of L w.r.t. the first argument is the constant 1 and therefore the squared loss is 1-smooth.

Absolute loss:. $L(a, y) = |a - y|$. The domain $\mathcal{Y}$ is again a bounded subset of $\mathbb{R}$. Now, L is not differentiable. However, L is 1-Lipschitz.

Properties of loss and distribution:				**#features needed to guarantee** $R(\mathbf{w}) - R(\bar{\mathbf{w}}) \leq \epsilon$
L is ρ-Lipschitz	L is β-smooth	R is λ-strongly convex on $\text{supp}(\bar{\mathbf{w}})$	R is sparsely λ-strongly convex	
X				$\|\bar{\mathbf{w}}\|_1^2 \, \frac{\rho^2}{\epsilon^2}$
	X			$\|\bar{\mathbf{w}}\|_1^2 \, \frac{\beta}{\epsilon}$
X		X		$\|\bar{\mathbf{w}}\|_0 \, \frac{1}{\lambda \, \epsilon}$
	X	X		$\|\bar{\mathbf{w}}\|_0 \, \frac{\beta}{\lambda \, \epsilon}$
	X	X	X	$\|\bar{\mathbf{w}}\|_0 \, \min\left\{ \frac{\beta}{\lambda} \log\left(\frac{1}{\epsilon}\right) \, , \, 1 + \frac{4\beta^2}{\lambda^2} \right\}$

TABLE 2.1
Summary of Results

Logistic-loss:. $L(a, y) = \log(1 + \exp(-y\, a))$. The domain $\mathcal{Y}$ is $\{+1, -1\}$. The derivative of L w.r.t. the first argument is the function $L'(a, y) = \frac{-y}{1+\exp(y\, a)}$. Since $L'(a, y) \in [-1, 1]$ we get that L is 1-Lipschitz. In addition, the second derivative of L is $-y \frac{1}{1+\exp(y\, a)} \frac{1}{1+\exp(-y\, a)} \in [-\frac{1}{4}, \frac{1}{4}]$ and therefore L is $\frac{1}{4}$ smooth.

Hinge-loss:. $L(a, y) = \max\{0, 1 - y\, a\}$. The domain $\mathcal{Y}$ is $\{+1, -1\}$. Like the absolute loss, the hinge-loss is not differentiable but is 1 Lipschitz.

Theorem 2.1 implies that without making any additional assumption on the distribution over $\mathcal{X} \times \mathcal{Y}$, for any B_1 and ϵ we can compute a predictor $\mathbf{w}$ such that

$$R(\mathbf{w}) \leq \min_{\mathbf{w}' : \|\mathbf{w}'\|_1 \leq B_1} R(\mathbf{w}') + \epsilon$$

and $\|\mathbf{w}\|_0 \leq \frac{8\sqrt{2}\, B_1^2}{\epsilon^2}$ for absolute-loss and hinge-loss, $\|\mathbf{w}\|_0 \leq \frac{B_1^2}{\epsilon}$ for squared-loss, and $\|\mathbf{w}\|_0 \leq \frac{B_1^2}{4\, \epsilon}$ for logistic-loss.

Next, we discuss possible applications of Theorems 2.8-2.9. Let L be the squared-loss function and assume that $\mathcal{Y} = [+1, -1]$. Therefore, for any $\bar{\mathbf{w}}$ we have $R(\mathbf{0}) - R(\bar{\mathbf{w}}) \leq 1$. We can rewrite $R(\mathbf{w})$ as

$$\begin{aligned} R(\mathbf{w}) &= \frac{1}{2}\, \mathbb{E}[(\langle \mathbf{w}, \mathbf{x} \rangle - y)^2] \\ &= \frac{1}{2} \mathbf{w}^T\, \mathbb{E}[\mathbf{x}\, \mathbf{x}^T]\, \mathbf{w} + \langle \mathbf{w}, \mathbb{E}[y\, \mathbf{x}] \rangle + \frac{1}{2} \mathbb{E}[y^2] \ . \end{aligned}$$

Thus, $R(\mathbf{w})$ is a quadratic function of $\mathbf{w}$ and therefore is λ-strongly convex where λ is the minimal eigenvalue of the matrix $\mathbb{E}[\mathbf{x}\, \mathbf{x}^T]$. Assuming that the instances $\mathbf{x}$ are uniformly distributed over $\{+1, -1\}^d$, then the features of $\mathbf{x}$ are uncorrelated, have zero mean, and have a unit variance. Therefore, the matrix $\mathbb{E}[\mathbf{x}\, \mathbf{x}^T]$ is the identity matrix and thus R is 1-strongly convex. Applying Theorem 2.8 we obtain that for any

$\bar{\mathbf{w}}$ we can efficiently find $\mathbf{w}$ such that $R(\mathbf{w}) \leq R(\bar{\mathbf{w}}) + \epsilon$ and $\|\mathbf{w}\|_0 \leq 2\,\|\bar{\mathbf{w}}\|_0 \log(1/\epsilon)$. Furthermore, applying Theorem 2.9 we obtain that one can find $\mathbf{w}$ such that $R(\mathbf{w})$ is only slightly larger than $R(\bar{\mathbf{w}})$ and $\|\mathbf{w}\|_0 \leq 5\,\|\bar{\mathbf{w}}\|_0$.

The argument above relies on the assumption that we fully know the conditional probability of the target y. This is a rather unrealistic assumption. It is more reasonable to assume that we have an i.i.d. sample of n examples from the distribution over $\mathcal{X} \times \mathcal{Y}$, where $n \ll d$, and let us redefine R using the uniform distribution over this sample. Now, $R(\mathbf{w})$ is no longer strongly convex, as the rank of the matrix $\mathbb{E}[\mathbf{x}\,\mathbf{x}^T]$ is n, while the dimension of the matrix is $d \gg n$. However, with high probability over the choice of the n examples, $R(\mathbf{w})$ is (k, λ)-sparsely-strongly convex with $n = O(k \ln d)$ and $\lambda = \frac{1}{2}$. This condition is often referred to as RIP (restricted isometry property) in the compressed sensing literature (4), which follows from concentration results in the random matrix literature. Therefore, we can still apply Theorem 2.8 and get that for any $\bar{\mathbf{w}}$ and k, such that $k \geq 2\,\|\bar{\mathbf{w}}\|_0 \log(1/\epsilon)$, we can efficiently find $\mathbf{w}$ such that $R(\mathbf{w}) - R(\bar{\mathbf{w}}) \leq \epsilon$ and $\|\mathbf{w}\|_0 \leq k$.

The strong convexity assumption given in Theorems 2.8-2.9 is much stronger than the one given in Corollary 2.3. To see this, note that the condition given in Theorem 2.8 breaks down even if we merely duplicate a single feature, while the condition of Corollary 2.3 is not affected by duplication of features. In fact, the condition of Corollary 2.3 still holds even if we construct many new features from the original features as long as $\bar{\mathbf{w}}$ will not change. Of course, the price we pay for relying on a much weaker assumption is an exponentially worse dependence on $1/\epsilon$.

Finally, as mentioned at the end of the previous section, the guarantees of Theorems 2.8-2.9 hold even if we add to $R(\mathbf{w})$ an ℓ_2 regularization term. For example, it holds for the problem of ℓ_2 regularized logistic regression:

$$R(\mathbf{w}) \;=\; \mathbb{E}[\log(1 + \exp(-y\,\langle \mathbf{w}, \mathbf{x}\rangle))] + \tfrac{\lambda}{2}\|\mathbf{w}\|_2^2 \;.$$

Since now $R(\mathbf{w})$ is everywhere strongly convex, we get that for any $\bar{\mathbf{w}}$ we can efficiently find $\mathbf{w}$ such that $R(\mathbf{w}) \leq R(\bar{\mathbf{w}}) + \epsilon$ and $\|\mathbf{w}\|_0 \leq O(\|\bar{\mathbf{w}}\|_0)$ in time $O(\log(1/\epsilon))$. Here, it is important to emphasize that the minimizer of $R(\mathbf{w})$ will in general be dense, since the ℓ_2 regularization tends to spread weights on many features. However, Theorems 2.8-2.9 hold for *any* reference vector, and not only for the minimizer of $R(\mathbf{w})$. It is therefore suffices that there is a sparse vector which gives a reasonable approximation to $R(\mathbf{w})$, and the theorem tells us that we will be competitive with this vector. We thus have two ways to control the trade-off between accuracy (i.e. low risk) and sparsity. One way is through the parameter ϵ. The second way is by choosing the reference vector $\bar{\mathbf{w}}$. We can have a sparser reference vector, over a non-correlated set of features, but this can also lead to a reference vector with higher risk.

4. Tightness. In this section we argue that some of the relations between sparsity, accuracy, and the ℓ_1 norm, derived in Section 2 are tight, and better guarantees cannot be obtained without adding more assumptions. This means that the procedures of Section 2 are optimal in the sense that no other procedure can yield a better sparsity guarantees (better by more than a constant factor).

The following two theorems establish the tightness of the bounds given in Theorem 2.1.

THEOREM 4.1. *For any $B_1 > 2$ and $l > 0$, there exists a data distribution, such that a (dense) predictor $\mathbf{w}$ with $\|\mathbf{w}\|_1 = B_1$ can achieve mean absolute-error ($L(a,b) = |a-b|$) less than l, but for any $\epsilon \leq 0.1$, at least $B_1^2/(45\,\epsilon^2)$ features must be used for achieving mean absolute-error less than ϵ.*

THEOREM 4.2. *For any $B_1 > 2$ and $l > 0$, there exists a data distribution, such that a (dense) predictor $\mathbf{w}$ with $\|\mathbf{w}\|_1 = B_1$ can achieve mean squared-error ($L(a,b) = (a-b)^2$) less than l, but for any $\epsilon \leq 0.1$, at least $B_1^2/(8\,\epsilon)$ features must be used for achieving mean squared-error less than ϵ.*

5. Related Work. The problem of finding sparse solutions to optimization problems using greedy methods or using the ℓ_1-norm was extensively studied in different disciplines. Some of the results we derive are variants of known techniques. Below we review related work, focusing on the contribution of this paper.

5.1. Randomized sparsification. As mentioned before, the randomized sparsification procedure (Algorithm 1) is not new and dates back to Maurey (22). It was also previously proposed by (24), as a tool for obtaining generalization bounds for AdaBoost (but their bound also depends on $\log(m)$, where m is the number of examples in the input distribution). Studying neural networks with bounded fan-in, (16) provided an upper bound similar to Theorem 2.4, for the special case of the squared-error loss. See also (14; 2; 27; 6). These authors also derived results similar to our result in Theorem 2.1 (although, with less generality). To the best of our knowledge the dependence on ℓ_0 we obtain in Corollary 2.3 is new.

5.2. Forward greedy selection. Forward greedy selection procedures for minimizing a convex objective subject to a polyhedron constraint date back to Frank-Wolfe algorithm (10). In fact, our Algorithm 2 can be viewed as a variant of Frank-Wolfe algorithm in which the step size is determined using a closed-form. Similar algorithms were proposed in (32; 6). Again, our data-dependent closed-form expression for determining the step size is different than the line search given in Algorithm 1.1 of (6). We note that (6) also gives the step-size $2/(t+3)$, but this step size is not data-dependent. The bound we derive in Theorem 2.4 is similar to Theorem 2.2 of (6), and similar analysis can be also found in (32) and even in the original analysis of Wolfe. To the best of our knowledge, the bound we derive for non-smooth but Lipschitz functions, based on the approximation given in Lemma 2.5 is novel. Additionally, as for the randomized sparsification procedure, we can obtain results with respect to the ℓ_0 norm, with the additional strong convexity requirement on the support of $\bar{\mathbf{w}}$, using Lemma 2.2.

Our Fully Corrective algorithm (Algorithm 3) is very similar to Algorithm 4.2 in (6), with one major difference —our algorithm do not enforce the constraint $\|\mathbf{w}\|_1 \leq B_1$ at all. This stands in contrast to many variants of the Fully Corrective algorithm studied in the context of (the dual of) the Minimum Enclosing Ball (MEB) problem by several authors (see e.g. (1; 6) and the references therein). This difference stems from the fact that our goal is *not* to solve the minimization problem with an ℓ_1 constraint but rather to solve the minimization problem with an ℓ_0 constraint. As discussed previously, a major advantage of not enforcing the constraint $\|\mathbf{w}\|_1 \leq B_1$ is that we are not required to tune the parameter B_1 in advance. Another important advantage is that we can derive results with respect to an arbitrary competing vector, $\bar{\mathbf{w}}$, which can be quite different from the minimizer of the convex optimization problem with the ℓ_1 constraint. See more discussion about the improtance of this issue in the next subsection.

The sparsity bound we derive in Theorem 2.7 depends logarithmically on $1/\epsilon$. A linear convergence result for a modified Frank-Wolfe algorithm with additional "away" steps (similar to our replacement steps) was derived by (13). However, their bound requires strong-convexity while our bound only requires sparsely-strong-convexity.

Furthermore, we derive convergence rates with respect to the ℓ_0 norm of the competing vector while the result of (13) only deals with convergence to the minimizer of the problem with respect to the ℓ_1 constraint. Such an analysis is clearly not satisfactory in our case. In the context of solving the dual of the MEB problem, (1) proves linear convergence for Frank-Wolfe with away steps under milder conditions than the strong-convexity assumption of (13). It seems that the result proved by (1) gives a local linear convergence rate, but unfortunately not an improved global complexity bound. More importantly, the results of (1), like the convergence analysis of (13), only deals with convergence to the solution of the convex problem (with the ℓ_1 constraint). In contrast, our analysis is with respect to the non-convex problem of minimizing the objective with an ℓ_0 constraint.

The post-processing step we perform (Algorithm 4) is in some sense similar to the idea of the modified Wolfe algorithm with the additional "away" steps —see for example (13; 1). It is also similar to Algorithm 5.1 in (6) and is also employed in (15). As before, a major difference is that we do not aim at solving the convex optimization problem with the ℓ_1 constraint and therefore do not enforce this constraint at all. As discussed previously, the advantages of not enforcing this constraint are that we do not need to tune the parameter B_1 and we can give guarantees with respect to an arbitrary competing vector. Additionally, the result we derive in Theorem 2.8 tells us that we can approximate $R(\bar{\mathbf{w}})$ to arbitrary precision using a number of features which is at most a *constant* approximation factor. Clearly, such a result cannot be obtained from the convergence analysis given in (13; 1; 6; 15).

Finally, we mention that forward greedy selection algorithms are called boosting in the machine learning literature (see e.g. (11)). It was observed empirically that Fully Corrective algorithms are usually more efficient than their corresponding non-corrective versions (see e.g. (31; 30)). In this paper we give a partial theoretical explanation to this empirical observation. For regression with the quadratic loss, this method is referred to as matching pursuit in the signal processing community (18).

5.3. Sparsistency of the Lasso. The use of the ℓ_1-norm as a surrogate for sparsity has a long history (e.g. (28) and the references therein), and much work has been done on understanding the relationship between the ℓ_1-norm and sparsity. Studying sparsity properties of the "Lasso", and feature selection techniques, several recent papers establish exact recovery of a sparse predictor based on the ℓ_1 relaxation (e.g. (34) and the references therein). One of the strongest result is the one derived for the FoBa algorithm of (33). However, for exact recovery much stronger conditions are required. In particular, *all* results that establish exact recovery require that the data will be generated (at least approximately) by a sparse predictor. In contrast, our bounds hold with respect to *any* reference predictor $\bar{\mathbf{w}}$. In practical applications, such as medical diagnosis, this is a big difference. For example, if the task is to predict illness using medical tests, it is very natural to assume that there exists a very sparse predictor with error of say 0.1, while a very dense predictor is required to achieve error below 0.05. In this case, exact recovery of a sparse predictor is impossible (because the best predictor is dense), but one can still compromise on the accuracy and achieve a very sparse predictor with a reasonable level of accuracy. Another requirement for exact recovery is that the magnitude of any non-zero element of $\mathbf{w}$ is large. We do not have such a requirement. Finally, all exact recovery results require the sparse eigenvalue condition. This condition is often referred to as RIP (restricted isometry property) in the compressed sensing literature (4). In contrast, as discussed in previous sections, some of our results require much weaker conditions. This is

attributed to the fact that our goal is different – we do not care about finding $\mathbf{w}^\star$ exactly but solely concern about finding some $\mathbf{w}$, with a good balance of low risk and sparsity. By compromising on accuracy, we get sparsity guarantees under much milder conditions.

5.4. Compressed sensing. As mentioned previously, recent work on compressed sensing (4; 5; 8; 9) also provide sufficient conditions for when the minimizer of the ℓ_1 relaxed problem is also the solution of the ℓ_0 problem. But, again, the assumptions are much stronger. We note that in compressed sensing applications, we have a control on the distribution over $\mathcal{X}$ (i.e. the design matrix). Therefore, the sparse eigenvalue condition (equivalently, RIP) is under our control. In contrast, in machine learning problems the distribution over $\mathcal{X}$ is provided by nature, and RIP conditions usually do not hold.

5.5. Learning theory. ℓ_1 norm have also been studies in learning theory as a regularization technique. For example, (17) showed that multiplicative online learning algorithms can be competitive with a sparse predictor, even when there are many irrelevant features, while additive algorithms are likely to make much more errors. This was later explained by the fact that multiplicative algorithms can be derived from an entropy regularization, which is strongly convex with respect to the ℓ_1 norm, while additive algorithms are derived from an ℓ_2 regularization (see e.g. (25)). Similarly, (20) considered PAC learning of a sparse predictor, and showed 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. The fact that we learn a sparse predictor can be used to derive generalization bounds as well (for example, as in (24)). However, if we are only interested in prediction performance and generalization bounds, it is not necessarily true that sparsity is the best method for obtaining good generalization properties.

6. Discussion and future work. We described and analyzed a few efficient methods for sparsity constrained optimization problems encountered in statistics and machine learning. The sparsity bounds we obtain depend on the accuracy of the computed predictor. They also depend on the ℓ_1 norm of a reference predictor either explicitly (Theorems 2.1, 2.4, 2.7) or implicitly by imposing a strong convexity assumption and bounding the ℓ_1 norm of the reference vector using its ℓ_0 norm (Corollary 2.3 and Theorems 2.8-2.9). In all cases, the trade-off between sparsity and accuracy is controlled by the excess loss allowed (ϵ) and by choosing a reference vector with low ℓ_1 norm.

As we have shown in Section 4, some of the sparseness bounds we derived are tight, in the sense that there exists a distribution for which the relation between $\|\mathbf{w}\|_0$, $\|\bar{\mathbf{w}}\|_1$, and ϵ cannot be improved.

There are several possible extensions to this work. First, our Fully Corrective greedy selection algorithms assume that the domain of R is the entire Euclidean space. In some cases it is desirable to impose additional convex constraints on $\mathbf{w}$. We believe that our proof technique can be generalized to include simple constraints, such as box constraints. Another interesting direction is to further quantify the advantage of Fully Corrective methods over non-fully corrective methods.

Currently, our technique for obtaining bounds that involve the ℓ_0 of $\bar{\mathbf{w}}$ assumes

that the risk R is strongly convex on the support of $\bar{\mathbf{w}}$. While this condition is reasonable in the case of regression problems with the squared loss, it is less likely to hold in classification problems, when other loss functions are used. Developing alternative techniques for obtaining bounds that involve the ℓ_0 norm of $\bar{\mathbf{w}}$ in binary classification problems is therefore a challenging task.

References.

[1] S.D. Ahipasaoglu, Sun P., and Todd M.J., *Linear convergence of a modified frank-wolfe algorithm for computing minimum volume enclosing ellipsoids*, Optimization Methods and Software, 23 (2008), pp. 5–19.

[2] Andrew R. Barron, *Universal approximation bounds for superposition of a sigmoidal function*, IEEE Transactions on Information Theory, 39 (1993), pp. 930–945.

[3] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization*, Springer, 2006.

[4] E.J. Candes and T. Tao, *Decoding by linear programming*, IEEE Trans. on Information Theory, 51 (2005), pp. 4203–4215.

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

[6] K.L. Clarkson, *Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm*, in Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, 2008, pp. 922–931.

[7] G. Davis, S. Mallat, and M. Avellaneda, *Greedy adaptive approximation*, Journal of Constructive Approximation, 13 (1997), pp. 57–98.

[8] D.L. Donoho, *Compressed sensing*, in Technical Report, Stanford University, 2006.

[9] ——, *For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution*, in Comm. Pure Appl. Math. 59, 2006.

[10] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[11] Y. Freund and R. E. Schapire, *A short introduction to boosting*, Journal of Japanese Society for Artificial Intelligence, 14 (1999), pp. 771–780.

[12] J.H. Friedman, *Greedy function approximation: A gradient boosting machine*, Annals of Statistics, 29 (2001), pp. 1189–1232.

[13] J Guélat and P Marcotte, *Some comments of wolfe's 'away step'*, Math. Program., 35 (1986), pp. 110–119.

[14] Lee K. Jones, *A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training*, Annals of Statistics, 20 (1992), pp. 608–613.

[15] Piyush Kumar and E. Alper Yildirim, *An Algorithm and a Core Set Result for the Weighted Euclidean One-Center Problem*, INFORMS JOURNAL ON COMPUTING, 21 (2009), pp. 614–629.

[16] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson, *Efficient agnostic learning of neural networks with bounded fan-in*, IEEE Transactions on Information Theory, 42 (1996), pp. 2118–2132.

[17] N. Littlestone, *Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm*, Machine Learning, 2 (1988), pp. 285–318.

[18] S. Mallat and Z. Zhang, *Matching pursuits with time-frequency dictionaries*, IEEE Transactions on Signal Processing, 41 (1993), pp. 3397–3415.

[19] B. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Computing, 25 (1995), pp. 227–234.
[20] A.Y. Ng, *Feature selection, l_1 vs. l_2 regularization, and rotational invariance*, in Proceedings of the Twenty-First International Conference on Machine Learning, 2004.
[21] K. Oleszkiewicz, *On a nonsymmetric version of the khinchine-kahane inequality*, Progress In Probability, 56 (2003).
[22] G. Pisier, *Remarques sur un résultat non publié de B. maurey*, 1980-1981.
[23] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
[24] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee, *Boosting the margin: A new explanation for the effectiveness of voting methods*, in Machine Learning: Proceedings of the Fourteenth International Conference, 1997, pp. 322–330. To appear, *The Annals of Statistics*.
[25] S. Shalev-Shwartz, *Online Learning: Theory, Algorithms, and Applications*, PhD thesis, The Hebrew University, 2007.
[26] S. Shalev-Shwartz and Y. Singer, *On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms*, in Proceedings of the Nineteenth Annual Conference on Computational Learning Theory, 2008.
[27] S. Shalev-Shwartz and N. Srebro, *Low ℓ_1 norm and guarantees on sparsifiability*, in Sparse Optimization and Variable Selection, Joint ICML/COLT/UAI Workshop, 2008.
[28] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal. Statist. Soc B., 58 (1996), pp. 267–288.
[29] V. N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.
[30] M. Warmuth, K. Glocer, and S.V.N. Vishwanathan, *Entropy regularized lpboost*, in Algorithmic Learning Theory (ALT), 2008.
[31] M. Warmuth, J. Liao, and G. Ratsch, *Totally corrective boosting algorithms that maximize the margin*, in Proceedings of the 23rd international conference on Machine learning, 2006.
[32] T. Zhang, *Sequential greedy approximation for certain convex optimization problems*, IEEE Transaction on Information Theory, 49 (2003), pp. 682–691.
[33] Tong Zhang, *Adaptive forward-backward greedy algorithm for sparse learning with linear models*, in NIPS, 2008.
[34] P. Zhao and B. Yu, *On model selection consistency of Lasso*, Journal of Machine Learning Research, 7 (2006), pp. 2541–2567.

Appendix A. Proofs.

A.1. Proof of Theorem 2.1. Without loss of generality, we assume that $\mathbf{w}_i^\star \geq 0$ for all i. Let $\mathbf{r} = (r_1, \ldots, r_k)$ be the sequence of random indices the randomized sparsification procedure chooses, and let $\mathbf{w}$ be the output of the procedure. Note that $\mathbf{w}$ is a function of $\mathbf{r}$ and therefore it is a random variable.

Let $\mathbf{x}$ be a given vector. Then, it is easy to verify that

$$\mathbb{E}_{\mathbf{r}}[\langle \mathbf{w}, \mathbf{x} \rangle] = \langle \mathbf{w}^\star, \mathbf{x} \rangle \ . \tag{A.1}$$

In the following, we first analyze the expected value of $R(\mathbf{w}) - R(\mathbf{w}^\star)$ for the two possible assumptions on L.

Lemma A.1. *Assume that the conditions of Theorem 2.1 hold and that L is*

β-smooth. Then:

$$\mathbb{E}_{\mathbf{r}}\left[R(\mathbf{w}) - R(\mathbf{w}^\star)\right] \;\leq\; \frac{\beta\,\|\mathbf{w}^\star\|_1^2}{2\,k}\;.$$

Proof. Since L is β-smooth, we can use the first inequality in Lemma B.1 to get that

$$R(\mathbf{w}) - R(\mathbf{w}^\star) \leq \langle \nabla R(\mathbf{w}^\star), \mathbf{w} - \mathbf{w}^\star\rangle + \frac{\beta}{2}\mathbb{E}_{\mathbf{x}}[(\langle \mathbf{w} - \mathbf{w}^\star, \mathbf{x}\rangle)^2]\;.$$

Taking expectation over $\mathbf{r}$ and using Equation (A.1) we get

$$\begin{aligned}\mathbb{E}_{\mathbf{r}}[R(\mathbf{w}) - R(\mathbf{w}^\star)] \;&\leq\; \frac{\beta}{2}\mathbb{E}_{\mathbf{r},\mathbf{x}}[(\langle \mathbf{w} - \mathbf{w}^\star, \mathbf{x}\rangle)^2] \\ &=\; \frac{\beta}{2}\mathbb{E}_{\mathbf{x},\mathbf{r}}[(\langle \mathbf{w} - \mathbf{w}^\star, \mathbf{x}\rangle)^2]\;,\end{aligned}$$

where in the last equality we used the linearity of expectation. Next, we note that for any $\mathbf{x}$ the expression $\mathbb{E}_{\mathbf{r}}\left[(\langle \mathbf{w} - \mathbf{w}^\star, \mathbf{x}\rangle)^2\right]$ is the variance of the random variable $\langle \mathbf{w}, \mathbf{x}\rangle = \frac{\|\mathbf{w}^\star\|_1}{k}\sum_{i=1}^k x_{r_i}$. Since each random variable x_{r_i} is in $[-1, +1]$, its variance is at most 1. Therefore, using the fact that the random variables are independent, we obtain that the variance of $\langle \mathbf{w}, \mathbf{x}\rangle$ is at most $\frac{\|\mathbf{w}^\star\|_1^2}{k}$. This holds for any $\mathbf{x}$ and therefore also for the expectation over $\mathbf{x}$, and this concludes our proof. □

Next, we deal with the case of Lipschitz loss function.

LEMMA A.2. *Assume that the conditions of Theorem 2.1 hold and that L is ρ-Lipschitz. Then:*

$$\mathbb{E}_{\mathbf{r}}\left[R(\mathbf{w}) - R(\mathbf{w}^\star)\right] \;\leq\; \frac{\rho\,\|\mathbf{w}^\star\|_1}{\sqrt{k}}\;.$$

Proof. Since L is ρ-Lipschitz, we have for all $(\mathbf{x}, y)$

$$L(\langle \mathbf{w}, \mathbf{x}\rangle, y) - L(\langle \mathbf{w}^\star, \mathbf{x}\rangle, y) \;\leq\; \rho\,|\langle \mathbf{w}, \mathbf{x}\rangle - \langle \mathbf{w}^\star, \mathbf{x}\rangle|\;.$$

Taking expectation over $\mathbf{r}$ and $(\mathbf{x}, y)$ we get

$$\begin{aligned}\mathbb{E}_{\mathbf{r}}[R(\mathbf{w})] - R(\mathbf{w}^\star) \;&\leq\; \rho\,\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{r}}[|\langle \mathbf{w}, \mathbf{x}\rangle - \langle \mathbf{w}^\star, \mathbf{x}\rangle|] \\ &\leq\; \rho\,\sqrt{\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{r}}[|\langle \mathbf{w}, \mathbf{x}\rangle - \langle \mathbf{w}^\star, \mathbf{x}\rangle|^2]}\;,\end{aligned}$$

where the last inequality follows from Jensen's inequality. The same argument as in the previous lemma concludes our proof. □

Equipped with the above we are now ready to prove Theorem 2.1. First, since $\mathbf{w}^\star$ is a minimizer of R over the ℓ_1 ball of radius $\|\mathbf{w}^\star\|_1$ and since $\mathbf{w}$ is in this ball, we obtain that $R(\mathbf{w}) - R(\mathbf{w}^\star)$ is a non-negative random variable. Therefore, using Markov's inequality, we get that with probability of at least 0.5 we have

$$R(\mathbf{w}) - R(\mathbf{w}^\star) \;\leq\; 2\mathbb{E}_r[R(\mathbf{w}) - R(\mathbf{w}^\star)]\;.$$

Plugging the bounds on $\mathbb{E}_r[R(\mathbf{w}) - R(\mathbf{w}^\star)]$ from the previous two lemmas, letting the right-hand side be ϵ and solving for k we conclude our proof.

A.2. Proof of Lemma 2.2. Since $\text{supp}(\bar{\mathbf{w}}) = F$ and since $\bar{\mathbf{w}}$ is optimal over the features in F we have that $\langle \nabla R(\bar{\mathbf{w}}), \bar{\mathbf{w}} \rangle = 0$. Therefore, using the assumption that R is λ-strongly convex on F we obtain

$$\begin{aligned} R(\mathbf{0}) - R(\bar{\mathbf{w}}) &= R(\mathbf{0}) - R(\bar{\mathbf{w}}) - \langle \nabla R(\bar{\mathbf{w}}), \mathbf{0} - \bar{\mathbf{w}} \rangle \\ &\geq \frac{\lambda}{2} \|\bar{\mathbf{w}} - \mathbf{0}\|_2^2 \end{aligned}$$

which implies that

$$\|\bar{\mathbf{w}}\|_2^2 \;\leq\; \frac{2\left(R(\mathbf{0}) - R(\bar{\mathbf{w}})\right)}{\lambda} \;.$$

Finally, we use the fact that $\bar{\mathbf{w}}$ has effective dimension of $\|\bar{\mathbf{w}}\|_0$ to get that

$$\|\bar{\mathbf{w}}\|_1^2 \leq \|\bar{\mathbf{w}}\|_0 \, \|\bar{\mathbf{w}}\|_2^2 .$$

Combining the above inequalities we conclude our proof.

A.3. Proof of Theorem 2.4. For all t, let $\epsilon_t = R(\mathbf{w}^{(t)}) - R(\mathbf{w}^\star)$ be the sub-optimality of the algorithm at iteration t. The following lemma provides us with an upper bound on ϵ_t. Its proof uses duality arguments (see for example (23; 3)).

LEMMA A.3. $\langle \boldsymbol{\theta}^{(k)}, \mathbf{w}^{(k)} \rangle + B_1 \, \|\boldsymbol{\theta}^{(k)}\|_\infty \;\geq\; \epsilon_k$.

Proof. We denote the Fenchel conjugate of R by $R^\star$. The Fenchel dual problem of Equation (2.1) is to maximize over $\boldsymbol{\theta} \in \mathbb{R}^d$ the objective $-R^\star(\boldsymbol{\theta}) - B_1 \, \|\boldsymbol{\theta}\|_\infty$. Therefore, the weak duality theorem tells us that for any $\boldsymbol{\theta}$

$$-R^\star(\boldsymbol{\theta}) - B_1 \, \|\boldsymbol{\theta}\|_\infty \;\leq\; R(\mathbf{w}^\star) \;\leq\; R(\mathbf{w}^{(k)}) \;.$$

Thus,

$$\epsilon_k \;\leq\; R(\mathbf{w}^{(k)}) + R^\star(\boldsymbol{\theta}) + B_1 \, \|\boldsymbol{\theta}\|_\infty \;. \tag{A.2}$$

In particular, it holds for $\boldsymbol{\theta}^{(k)} = \nabla R(\mathbf{w}^{(k)})$. Next, we use (3, Proposition 3.3.4) to get that for $\boldsymbol{\theta}^{(k)} = \nabla R(\mathbf{w}^{(k)})$ we have $R(\mathbf{w}^{(k)}) + R^\star(\boldsymbol{\theta}^{(k)}) = \langle \mathbf{w}^{(k)}, \boldsymbol{\theta}^{(k)} \rangle$. Combining this with Equation (A.2) we conclude our proof. □

The next lemma analyzes the progress of the algorithm.

LEMMA A.4. *The sequence $\epsilon_1, \epsilon_2, \ldots$ is monotonically non-increasing. Furthermore, let T be the minimal integer such that $\epsilon_T \leq 4\beta B_1^2$. Then, for $t < T$ we have $\epsilon_t - \epsilon_{t+1} \geq 2\,\beta\, B_1^2$ and for $t \geq T$ we have*

$$\epsilon_t - \epsilon_{t+1} \geq \epsilon_t^2 \, \frac{1}{8\,\beta\, B_1^2} \;\;.$$

Proof. To simplify the proof, we assume without loss of generality that $\text{sgn}(\theta_{r_t}^{(t)}) = -1$. Denote $\mathbf{u}^{(t)} = \eta_t(B_1 \mathbf{e}^{r_t} - \mathbf{w}^{(t)})$ and thus we can rewrite the update rule as $\mathbf{w}^{(t+1)} = (1 - \eta_t)\mathbf{w}^{(t)} + \eta_t \, B_1 \, \mathbf{e}^{r_t} = \mathbf{w}^{(t)} + \mathbf{u}^{(t)}$. Let $\Delta_t = \epsilon_t - \epsilon_{t+1} = R(\mathbf{w}^{(t)}) - R(\mathbf{w}^{(t+1)})$. Using the assumption that L is β-smooth and Lemma B.1 we obtain that

$$\Delta_t \;\geq\; -\langle \boldsymbol{\theta}^{(t)}, \mathbf{u}^{(t)} \rangle - \frac{\beta \, \|\mathbf{u}^{(t)}\|_1^2}{2} \;.$$

Next, we use the definition of $\mathbf{u}^{(t)}$, the triangle inequality, and the fact that $\|\mathbf{w}^{(t)}\|_1 \leq B_1$ to get that

$$\|\mathbf{u}^{(t)}\|_1 \leq \eta_t(\|B_1\mathbf{e}^{r_t}\|_1 + \|\mathbf{w}^{(t)}\|_1) \leq 2\,\eta_t\,B_1 \quad .$$

Therefore,

$$\begin{aligned}\Delta_t &\geq -\langle\boldsymbol{\theta}^{(t)}, \mathbf{u}^{(t)}\rangle - 2\,\beta\,\eta_t^2\,B_1^2 \\ &= \eta_t\left(\langle\boldsymbol{\theta}^{(t)}, \mathbf{w}^{(t)}\rangle - B_1\,\langle\boldsymbol{\theta}^{(t)}, \mathbf{e}^{r_t}\rangle\right) - 2\,\beta\,\eta_t^2\,B_1^2 \quad .\end{aligned} \tag{A.3}$$

The definition of r_t implies that $\langle\boldsymbol{\theta}^{(t)}, \mathbf{e}^{r_t}\rangle = -\|\boldsymbol{\theta}^{(t)}\|_\infty$. Therefore, we can invoke Lemma A.3 and obtain that $0 \leq \epsilon_t \leq \langle\boldsymbol{\theta}^{(t)}, \mathbf{w}^{(t)}\rangle - B_1\,\langle\boldsymbol{\theta}^{(t)}, \mathbf{e}^{r_t}\rangle$. Next, we note that η_t is defined to be the maximizer of the right-hand side of Equation (A.3) over $[0, 1]$. Therefore, for any $\eta \in [0, 1]$ we have

$$\begin{aligned}\Delta_t &\geq \eta\left(\langle\boldsymbol{\theta}^{(t)}, \mathbf{w}^{(t)}\rangle - B_1\,\langle\boldsymbol{\theta}^{(t)}, \mathbf{e}^{r_t}\rangle\right) - 2\,\beta\,\eta^2\,B_1^2 \\ &\geq \eta\,\epsilon_t - 2\,\beta\,\eta^2\,B_1^2\ .\end{aligned} \tag{A.4}$$

If $\epsilon_t \leq 4\beta B_1^2$ then by setting $\eta = \frac{\epsilon_t}{4\,\beta\,B_1^2}$ we obtain $\Delta_t \geq \frac{\epsilon_t^2}{8\,\beta\,B_1^2}$. If $\epsilon_t > 4\,\beta B_1^2$ then setting $\eta = 1$ gives $\Delta_t \geq 2\,\beta B_1^2$. □

We are now ready to prove Theorem 2.4. First, the inequality (A.4) with $\eta = 1$ and $t = 0$ implies that

$$\epsilon_0 - \epsilon_1 = \Delta_0 \geq \epsilon_0 - 2\beta B_1^2.$$

This means that $\epsilon_1 \leq 2\beta B_1^2$. Therefore starting from $t \geq 1$ we can apply the same argument of Lemma B.2 and this concludes our proof.

A.4. Proof of Lemma 2.5. For simplicity, we omit the second argument of L and $\tilde{L}$ throughout the proof. We first prove that $\tilde{L}$ is β-smooth. The proof uses ideas from convex analysis. We refer the reader to (23; 3) and see also a similar derivation in (26). The definition of $\tilde{L}$ implies that it is the infimal convolution of L and the quadratic function $(\beta/2)v^2$. Therefore, using the infimal convolution theorem (23, Chapter 16) we obtain that the Fenchel conjugate of $\tilde{L}$ is $\tilde{L}^\star(\theta) = \frac{1}{2\beta}\theta^2 + L^\star(\theta)$, where $L^\star$ is the Fenchel conjugate of L. Since the quadratic function is strongly convex we obtain that $\tilde{L}^\star$ is a $1/\beta$ strongly convex function, and thus its Fenchel conjugate, namely $\tilde{L}$, is β-smooth (25, Lemma 15). Next, we turn to the proof of $|L(a) - \tilde{L}(a)| \leq \frac{\rho^2}{2\,\beta}$. Let $f(v) = \frac{\beta}{2}v^2 + L(a - v)$. On one hand, $\tilde{L}(a) \leq f(0) = L(a)$. On the other hand, since $L(a) - L(a - v) \leq \rho\,|v|$ we have

$$f(v) = \frac{\beta}{2}v^2 + L(a) + L(a - v) - L(a) \geq \frac{\beta}{2}v^2 + L(a) - \rho\,|v|\ .$$

Therefore,

$$\tilde{L}(a) = \inf_v f(v) \geq L(a) + \inf_v\left[\frac{\beta}{2}v^2 - \rho\,v\right] = L(a) - \frac{\rho^2}{2\,\beta}\ .$$

This concludes our proof.

A.5. Proof of Theorem 2.7. We start with the following lemma which states that if the greedy algorithm has not yet identified all the features of $\bar{\mathbf{w}}$ then a single greedy iteration yields a substantial progress.

LEMMA A.5. *Let $F, \bar{F}$ be two subsets of $[d]$ such that $\bar{F} - F \neq \emptyset$ and let*

$$\mathbf{w} = \operatorname*{argmin}_{\mathbf{v}:\mathrm{supp}(\mathbf{v})=F} R(\mathbf{v}) \quad , \quad \bar{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{v}:\mathrm{supp}(\mathbf{v})=\bar{F}} R(\mathbf{v}) \ .$$

Assume that L is β-smooth and that

$$R(\bar{\mathbf{w}}) - R(\mathbf{w}) - \langle \nabla R(\mathbf{w}), \bar{\mathbf{w}} - \mathbf{w} \rangle \geq \frac{\lambda}{2} \|\mathbf{w} - \bar{\mathbf{w}}\|_2^2 \ . \tag{A.5}$$

Then,

$$R(\mathbf{w}) - \min_{\alpha} R(\mathbf{w} + \alpha \mathbf{e}^j) \ \geq \ \frac{\left(R(\mathbf{w}) - R(\bar{\mathbf{w}}) + \frac{\lambda}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|^2\right)^2}{2\,\beta\,\left(\sum_{i \in \bar{F} - F} |\bar{\mathbf{w}}_i|\right)^2},$$

where $j = \mathrm{argmax}_i \, |\nabla R(\mathbf{w}^{(k)})_i|$.

Proof. To simplify notation, denote $F^c = \bar{F} - F$. For all $j \in F^c$ and $\eta > 0$, we define

$$Q_j(\eta) = R(\mathbf{w}) + \eta \, \mathrm{sgn}(\bar{w}_j) \, \langle \nabla R(\mathbf{w}), \mathbf{e}^j \rangle + \frac{\eta^2 \, \beta}{2} \ .$$

Next, using the assumption that L is smooth and Lemma B.1 we obtain that

$$R(\mathbf{w} + \eta \, \mathrm{sgn}(\bar{w}_j) \, \mathbf{e}^j) \ \leq \ Q_j(\eta).$$

Since the choice of $j = \mathrm{argmax}_i \, |\nabla R(\mathbf{w}^{(k)})_i|$ achieves the minimum of $\min_j \min_\eta Q_j(\eta)$, the lemma is a direct consequence of the following stronger statement:

$$R(\mathbf{w}) - \min_{j} Q_j(\eta) \ \geq \ \frac{\left(R(\mathbf{w}) - R(\bar{\mathbf{w}}) + \frac{\lambda}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|^2\right)^2}{2\,\beta\,\left(\sum_{i \in \bar{F} - F} |\bar{\mathbf{w}}_i|\right)^2}, \tag{A.6}$$

for an appropriate choice of η. Therefore, we now turn to prove that Equation (A.6) holds.

Denote $s = \sum_{j \in F^c} |\bar{w}_j|$, we obtain that

$$\begin{aligned} s \, \min_j Q_j(\eta) \leq & \sum_{j \in F^c} |\bar{w}_j| Q_j(\eta) \\ & \leq \ s \, R(\mathbf{w}) + \eta \sum_{j \in F^c} \bar{w}_j \, (\nabla R(\mathbf{w}))_j + s \, \frac{\eta^2 \, \beta}{2} \ . \end{aligned} \tag{A.7}$$

Since we assume that $\mathbf{w}$ is optimal over F we get that $(\nabla R(\mathbf{w}))_j = 0$ for all $j \in F$. Additionally, $w_j = 0$ for $j \notin F$ and $\bar{w}_j = 0$ for $j \notin \bar{F}$. Therefore,

$$\begin{aligned} \sum_{j \in F^c} \bar{w}_j \, (\nabla R(\mathbf{w}))_j &= \sum_{j \in F^c} (\bar{w}_j - w_j) \, (\nabla R(\mathbf{w}))_j \\ &= \sum_{j \in \bar{F} \cup F} (\bar{w}_j - w_j) \, (\nabla R(\mathbf{w}))_j \\ &= \langle \nabla R(\mathbf{w}), \bar{\mathbf{w}} - \mathbf{w} \rangle \ . \end{aligned}$$

Combining the above with the assumption given in Equation (A.5) we obtain that

$$\sum_{j \in F^c} \bar{w}_j \left(\nabla R(\mathbf{w})\right)_j \leq R(\bar{\mathbf{w}}) - R(\mathbf{w}) - \tfrac{\lambda}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|_2^2 \ .$$

Combining the above with Equation (A.7) we get

$$\begin{aligned} s \, \min_j Q_j(\eta) \leq \ & s \, R(\mathbf{w}) + s \, \frac{\eta^2 \, \beta}{2} \\ & - \eta \, \left(R(\mathbf{w}) - R(\bar{\mathbf{w}}) + \frac{\lambda}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|^2 \right) \ . \end{aligned}$$

Setting $\eta = \left(R(\mathbf{w}) - R(\bar{\mathbf{w}}) + \frac{\lambda}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|^2\right) / (\beta \, s)$ and rearranging terms we conclude our proof of (A.6). □

Equipped with the above lemma we now turn to prove Theorem 2.7. Note that the lemma assumes that $R(\mathbf{w})$ is λ-strongly convex on the relevant support (Equation (A.5)). Since Theorem 2.7 does not make such an assumption, we will apply the lemma with $\lambda = 0$ (this merely requires that R is convex, which follows from our assumption that L is convex). The rest of the conditions stated in Lemma A.5 hold and therefore,

$$\begin{aligned} R(\mathbf{w}^{(k)}) - R(\mathbf{w}^{(k+1)}) \geq \ & \frac{\left(R(\mathbf{w}^{(k)}) - R(\bar{\mathbf{w}})\right)^2}{2 \, \beta \, \left(\sum_{i \in \bar{F} - F^{(k)}} |\bar{w}_i|\right)^2} \\ \geq \ & \frac{\left(R(\mathbf{w}^{(k)}) - R(\bar{\mathbf{w}})\right)^2}{2 \, \beta \, \|\bar{\mathbf{w}}\|_1^2} \ . \end{aligned}$$

Denote $\epsilon_k = R(\mathbf{w}^{(k)}) - R(\bar{\mathbf{w}})$ and note that the above implies that $\epsilon_{k+1} \leq \epsilon_k - \frac{\epsilon_k^2}{2 \, \beta \, \|\bar{\mathbf{w}}\|_1^2}$. Our proof is concluded by combining the above inequality with Lemma B.2.

A.6. Proof of Theorem 2.8. Denote $\epsilon_k = R(\mathbf{w}^{(k)}) - R(\bar{\mathbf{w}})$. The definition of the update implies that $R(\mathbf{w}^{(k+1)}) \leq \min_{i,\alpha} R(\mathbf{w}^{(k)} + \alpha \, \mathbf{e}^i)$. The conditions of Lemma A.5 hold and therefore we obtain that (with $F = F^{(k)}$)

$$\begin{aligned} \epsilon_k - \epsilon_{k+1} = R(\mathbf{w}^k) - R(\mathbf{w}^{(k+1)}) \geq \ & \frac{\left(\epsilon_k + \frac{\lambda}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|^2\right)^2}{2 \, \beta \, \left(\sum_{i \in \bar{F} - F} |\bar{\mathbf{w}}_i|\right)^2} \\ \geq \ & \frac{4 \epsilon_k \, \frac{\lambda}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|^2}{2 \, \beta \, \left(\sum_{i \in \bar{F} - F} |\bar{\mathbf{w}}_i|\right)^2} \geq \frac{\epsilon_k \, \sum_{i \in \bar{F} - F} |\bar{\mathbf{w}}_i|^2}{\frac{\beta}{\lambda} \, \left(\sum_{i \in \bar{F} - F} |\bar{\mathbf{w}}_i|\right)^2} \\ \geq \ & \frac{\epsilon_k}{\frac{\beta}{\lambda} \, |\bar{F} - F|} \geq \frac{\epsilon_k}{\frac{\beta}{\lambda} \, \|\bar{\mathbf{w}}\|_0} \ . \end{aligned} \tag{A.8}$$

Therefore, $\epsilon_{k+1} \leq \epsilon_k \left(1 - \frac{\lambda}{\beta \, \|\bar{\mathbf{w}}\|_0}\right)$. Applying this inequality recursively we obtain $\epsilon_{k+1} \leq \epsilon_0 \left(1 - \frac{\lambda}{\beta \, \|\bar{\mathbf{w}}\|_0}\right)^{k+1}$. Therefore, if $\epsilon_k \geq \epsilon$ we must have $\epsilon \leq \epsilon_0 \left(1 - \frac{\lambda}{\beta \, \|\bar{\mathbf{w}}\|_0}\right)^k$. Using the inequality $1 - x \leq \exp(-x)$ and rearranging terms we conclude that $k \leq \beta \, \|\bar{\mathbf{w}}\|_0 \, \log\left(\frac{\epsilon_0}{\epsilon}\right)$.

A.7. Proof of Theorem 2.9. We first prove the following lemma.

LEMMA A.6. *Let $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a convex β-smooth loss function and let $R(\mathbf{w})$ be as defined in Equation (1.1), where the expectation is w.r.t. an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose that the post-processing backward procedure (Algorithm 4) is run for t iterations with input $F^{(0)}$ and denote $|F^{(0)}| = k$. Let $\lambda > 0$ be a scalar, $\bar{\mathbf{w}} \in \mathbb{R}^d$ be a vector, such that*

$$k + 1 \;\geq\; \|\bar{\mathbf{w}}\|_0(1 + 4\beta^2/\lambda^2) \quad ,$$

and assume that R is $(k + 1 + \|\bar{\mathbf{w}}\|_0, \lambda)$-sparsely-strongly convex. Then,

$$R(\mathbf{w}^{(t)}) \leq R(\bar{\mathbf{w}}) + \min\left[\delta_t/\alpha, \Delta_0(1-\alpha)^t\right],$$

where $\alpha = \frac{2\,\beta}{\lambda(k+1-\|\bar{\mathbf{w}}\|_0)}$ and $\Delta_0 = \max(0, R(\mathbf{w}^{(0)}) - R(\bar{\mathbf{w}}))$.

Proof. We first analyze the effect of one replacement step. To simplify notation, we use the shorthand $\mathbf{w}$ instead of $\mathbf{w}^{(t)}$ and F instead of $F^{(t)}$. We also denote $\bar{F} = \mathrm{supp}(\bar{\mathbf{w}})$. Let $\tilde{F} = \bar{F} \cup F$, and let

$$\tilde{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}:\mathrm{supp}(\mathbf{w}) \subseteq \tilde{F}} R(\mathbf{w}).$$

Let $\bar{k} = \|\bar{\mathbf{w}}\|_0$. In a replacement step, we first perform a forward step. We can therefore apply the analysis of the Fully Corrective forward selection, and in particular, we obtain from Equation (A.8) with $\bar{\mathbf{w}}$ replaced by $\tilde{\mathbf{w}}$ that

$$R(\mathbf{w}) - R(\mathbf{w}') \geq \frac{(R(\mathbf{w}) - R(\tilde{\mathbf{w}}))}{\frac{\beta}{\lambda}\,|\tilde{F} - F|} \geq \frac{(R(\mathbf{w}) - R(\tilde{\mathbf{w}}))}{\frac{\beta}{\lambda}\,\bar{k}}. \tag{A.9}$$

Next, we remove the smallest element of $\mathbf{w}'$, denoted w'_q. Since $\mathbf{w}'$ minimizes the loss over F', and $q \in F'$, we have that $(\nabla R(\mathbf{w}'))_q = 0$. Therefore, from the β-smoothness of R we obtain

$$R(\mathbf{w}' - w'_q\mathbf{e}^q) - R(\mathbf{w}') \leq -w'_q\,(\nabla R(\mathbf{w}'))_q + \tfrac{\beta}{2}(w'_q)^2 = \tfrac{\beta}{2}(w'_q)^2 \ .$$

The definition of $\delta_t = R(\mathbf{w}) - R(\mathbf{w}' - w'_q\mathbf{e}^q)$ yields that the left-hand side of the above equals to $R(\mathbf{w}) - R(\mathbf{w}') - \delta_t$ and therefore we obtain that

$$\tfrac{\beta}{2}(w'_q)^2 \geq R(\mathbf{w}) - R(\mathbf{w}') - \delta_t \ . \tag{A.10}$$

Combining the above with Equation (A.9) gives that

$$(w'_q)^2 \geq \tfrac{2}{\beta}\left(\frac{(R(\mathbf{w}) - R(\tilde{\mathbf{w}}))}{\frac{\beta}{\lambda}\,\bar{k}} - \delta_t\right) \tag{A.11}$$

Next, we derive an upper bound on $(w'_q)^2$. We have

$$\begin{aligned}
(w'_q)^2 \;&\leq \sum_{j \in F - \bar{F}} (w'_j)^2/|F - \bar{F}| \\
&\leq \|\mathbf{w}' - \bar{\mathbf{w}}\|_2^2/(k + 1 - \bar{k}) \\
&\leq 2[\|\mathbf{w}' - \tilde{\mathbf{w}}\|_2^2 + \|\bar{\mathbf{w}} - \tilde{\mathbf{w}}\|_2^2]/(k + 1 - \bar{k}) \\
&\leq \frac{4[R(\mathbf{w}') + R(\bar{\mathbf{w}}) - 2R(\tilde{\mathbf{w}})]}{\lambda(k + 1 - \bar{k})}.
\end{aligned}$$

Comparing the above upper bound with the lower bound given in Equation (A.11) we obtain

$$\frac{2}{\beta}\left(\frac{(R(\mathbf{w})-R(\tilde{\mathbf{w}}))}{\frac{\beta}{\lambda}\,\bar{k}}-\delta_t\right) \;\leq\; \frac{4[R(\mathbf{w}')+R(\bar{\mathbf{w}})-2R(\tilde{\mathbf{w}})]}{\lambda(k+1-\bar{k})}.$$

To simplify notation, let $s=\frac{\lambda}{\beta\bar{k}}$ and recall that $\alpha=\frac{2\,\beta}{\lambda(k+1-\bar{k})}$. Rearranging the above inequality and using the definitions of s and α we obtain

$$\begin{aligned}\delta_t \;&\geq s\;(R(\mathbf{w})-R(\tilde{\mathbf{w}}))-\alpha\;(R(\mathbf{w}')+R(\bar{\mathbf{w}})-2R(\tilde{\mathbf{w}}))\\ &= s\;(R(\mathbf{w})-R(\bar{\mathbf{w}})+R(\bar{\mathbf{w}})-R(\tilde{\mathbf{w}}))\\ &\quad-\alpha\;(R(\mathbf{w}')-R(\bar{\mathbf{w}})+2\,(R(\bar{\mathbf{w}})-R(\tilde{\mathbf{w}})))\;.\end{aligned}$$

Next, using Equation (A.9) we know that

$$R(\mathbf{w}')\leq R(\mathbf{w})-s(R(\mathbf{w})-R(\tilde{\mathbf{w}}))\;.$$

Subtracting $R(\bar{\mathbf{w}})$ from both sides and using the fact that $R(\tilde{\mathbf{w}})\leq R(\bar{\mathbf{w}})$ we obtain that

$$R(\mathbf{w}')-R(\bar{\mathbf{w}})\leq R(\mathbf{w})-R(\bar{\mathbf{w}})-s(R(\mathbf{w})-R(\tilde{\mathbf{w}}))\leq(R(\mathbf{w})-R(\bar{\mathbf{w}}))(1-s)\;.$$

Thus,

$$\delta_t\geq(s-\alpha(1-s))\;(R(\mathbf{w})-R(\bar{\mathbf{w}}))+(s-2\,\alpha)\;(R(\bar{\mathbf{w}})-R(\tilde{\mathbf{w}}))\;. \tag{A.12}$$

Now, using simple algebraic manipulations and the assumption $k+1\geq\bar{k}(1+4\beta^2/\lambda^2)$ we obtain

$$s-2\,\alpha=\frac{\lambda^2(k+1-\bar{k})-4\beta^2\bar{k}}{\beta\bar{k}\lambda(k+1-\bar{k})}=\frac{\lambda^2(k+1)-\bar{k}(\lambda^2+4\beta^2)}{\beta\bar{k}\lambda(k+1-\bar{k})}\geq 0\;,$$

and

$$s-\alpha(1-s)\;=\;s-2\,\alpha+\alpha+\alpha\,s\geq\alpha\;.$$

Combine this with Equation (A.12) we get $\delta_t/\alpha\geq R(\mathbf{w})-R(\bar{\mathbf{w}})$. This proves the first half of the desired bound. Moreover, if we let $\Delta_t=R(\mathbf{w})-R(\bar{\mathbf{w}})$ at the beginning of the t-th iteration, then the inequality $\delta_t/\alpha\geq R(\mathbf{w})-R(\bar{\mathbf{w}})$ implies that

$$\frac{\Delta_t-\Delta_{t+1}}{\alpha}=\frac{R(\mathbf{w}^{(t)})-R(\mathbf{w}^{(t+1)})}{\alpha}\geq\frac{R(\mathbf{w}^{(t)})-R(\mathbf{w}'-w'_q\mathbf{e}^q)}{\alpha}=\frac{\delta_t}{\alpha}\geq\Delta_t\;.$$

Therefore, $\Delta_{t+1}\leq\Delta_t(1-\alpha)\leq\Delta_0(1-\alpha)^{t+1}$. This proves the second half of the desired bound. □

We can now easily prove Theorem 2.9. We have two cases. First, if the stopping condition is met then from the above lemma we obtain that $R(\mathbf{w}^{(t)})-R(\bar{\mathbf{w}})\leq\delta_t/\alpha\leq 0\leq\epsilon$. Second, if we perform t iterations without breaking, then we get

$$\epsilon\leq\Delta_0(1-\alpha)^t\leq\Delta_0 e^{-\alpha\,t}\leq(R(\mathbf{0})-R(\bar{\mathbf{w}}))e^{-\alpha\,t}\;.$$

Rearranging the above and using the definition of α concludes our proof.

A.8. Proofs of Theorem 4.1 and Theorem 4.2. Fix some $B_1 > 2$, $l > 0$, and $\epsilon < 0.1$. To prove the theorems, we present an input distribution $\mathcal{D}$, then demonstrate a specific (dense) predictor with $\|\mathbf{w}\|_1 = B_1$ and mean error l, and finally present a lower bound on mean error of any sparse predictor, from which we can conclude that any predictor $\mathbf{u}$ with mean error at most ϵ must satisfy $\|\mathbf{u}\|_0 \geq \Omega(B_1^2/(\epsilon^\alpha)$, with $\alpha = 1$ for squared-error and 2 for absolute-error.

The data distribution: Consider an instance space $\mathcal{X} = \{+1, -1\}^d$, 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 $\Pr[Y = 1] = \frac{1}{2}$. Next, the features $X_1, \ldots, X_n$ are identically distributed and are independent conditioned on Y, with $\Pr[X_i = y|Y = y] = \frac{1+a}{2}$, where $a = 1/B_1$. 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 $w_i = 1/(da)$ for all features i. To simplify our notation, we use the shorthand $\mathbb{E}\left[\langle \mathbf{w}, \mathbf{x}\rangle \mid y\right]$ for denoting $\mathbb{E}\left[\langle \mathbf{w}, \mathbf{x}\rangle \mid Y = y\right]$. Verifying that for both values of y we have

$$\begin{aligned} \mathbb{E}\left[\langle \mathbf{w}, X\rangle \mid y\right] &= d\frac{1}{da}\, a\, y = y \\ \mathrm{Var}[\langle \mathbf{w}, X\rangle \mid y] &= \tfrac{1-a^2}{d\,a^2} \end{aligned} \tag{A.13}$$

we immediately obtain that

$$\mathbb{E}[(\langle \mathbf{w}, X\rangle - Y)^2] = \frac{1-a^2}{d\,a^2} \leq \frac{1}{d\,a^2} \quad . \tag{A.14}$$

Additionally, using Jensen's inequality we obtain that:

$$\mathbb{E}[|\langle \mathbf{w}, X\rangle - Y|] \leq \sqrt{\mathbb{E}[(\langle \mathbf{w}, \mathcal{X}\rangle - Y)^2]} \leq \sqrt{\frac{1}{d\,a^2}}\ . \tag{A.15}$$

Recall that $a = 1/B_1$ and choose the dimension to be $d = B_1^2/l^\alpha$, where $\alpha = 1$ for the squared-error and $\alpha = 2$ for the absolute error. This implies that for both cases, $\|\mathbf{w}\|_1 = B_1$ and $R(\mathbf{w}) \leq l$.

Sparse prediction: Consider any predictor $\mathbf{u}$ with only B_0 non-zero coefficients. For such a predictor we have $\sum \mathbf{u}_i^2 \geq (\sum \mathbf{u}_i)^2/B_0$. Denote $\rho = \sum_i \mathbf{u}_i$. Fix some $y \in \{\pm 1\}$ and denote $\mu_y = \mathbb{E}[\langle \mathbf{u}, X\rangle|y]$. We have,

$$\mathbb{E}[\langle \mathbf{u}, X\rangle|y] \;=\; y\,a\,\rho \quad \text{and} \quad \mathrm{Var}\left[\langle \mathbf{u}, X\rangle \mid y\right] \;=\; (1-a^2)\|\mathbf{u}\|_2^2\ .$$

We start with the case of the squared-error.

$$\begin{aligned} \mathbb{E}[(\langle \mathbf{u}, X\rangle - y)^2|y] \;&=\; \mathrm{Var}\left[\langle \mathbf{u}, X\rangle \mid y\right] + (\mu_y - y)^2 \\ &= (1-a^2)\|\mathbf{u}\|_2^2 + (1-a\rho)^2 \\ &\geq (1-a^2)\rho^2/B_0 + (1-a\rho)^2\ . \end{aligned} \tag{A.16}$$

Thus,

$$\mathbb{E}[(\langle \mathbf{u}, X\rangle - Y)^2] \geq (1-a^2)\rho^2/B_0 + (1-a\rho)^2\ .$$

If $|\rho| < B_1/2$ then the right-hand side of the above is at least $1/4$. Otherwise,

$$\mathbb{E}[(\langle \mathbf{u}, X\rangle - Y)^2] \geq \frac{(1-a^2)B_1^2}{4B_0} = \frac{B_1^2 - 1}{4\,B_0}\ .$$

Since we assume $B_1 \geq 2$ we have $B_1^2 - 1 \geq B_1^2/2$ and we conclude that

$$\mathbb{E}[(\langle \mathbf{u}, X\rangle - Y)^2] \geq \min\left\{\frac{1}{4}, \frac{B_1^2}{8\,B_0}\right\} \ .$$

Thus, if we want that $R(\mathbf{w})$ will be at most ϵ we must have

$$\frac{B_1^2}{8\,B_0} \leq \epsilon \quad \Rightarrow \quad B_0 \geq \frac{B_1^2}{8\,\epsilon} \ ,$$

which concludes the proof of Theorem 4.2.

Next, we consider the case of the absolute-error (Theorem 4.1). Since we consider only $B_1 > 2$, we have $0.05 < 0.25 \leq \Pr[X_i = Y|y] \leq 0.75 < 0.95$, with the loss being an affine function (degree one polynomial) of X. We can therefore use Lemma B.3 to get that:

$$\mathbb{E}\left[\,|\langle \mathbf{u}, X\rangle - Y|\,|\,y\right] \ \geq \ 0.2\,\sqrt{\mathbb{E}\left[\,(\langle \mathbf{u}, X\rangle - Y)^2 \mid y\right]} \ .$$

Combining the above with Equation (A.16) we obtain that

$$\mathbb{E}[|\langle \mathbf{u}, X\rangle - Y|] \geq 0.2\sqrt{(1-a^2)\rho^2/B_0 + (1-a\rho)^2} \ .$$

The rest of the proof follows analogously to the case of squared-error.

Appendix B. Technical Lemmas.

LEMMA B.1. *Let $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a convex β-smooth loss function and let $R(\mathbf{w})$ be as defined in Equation (1.1), where the expectation is w.r.t. an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Then, for any vectors $\mathbf{w}, \mathbf{u}$ we have*

$$R(\mathbf{w}+\mathbf{u}) - R(\mathbf{w}) - \langle \nabla R(\mathbf{w}), \mathbf{u}\rangle \ \leq \ \frac{\beta}{2}\mathbb{E}_{\mathbf{x}}[(\langle \mathbf{u}, \mathbf{x}\rangle)^2] \ \leq \ \frac{\beta\,\|\mathbf{u}\|_1^2}{2} \ .$$

Proof. Since L is β-smooth we have for any $\mathbf{w}, \mathbf{u}$ and $(\mathbf{x}, y)$, $L(\langle \mathbf{w}+\mathbf{u}, \mathbf{x}\rangle, y) - L(\langle \mathbf{w}, \mathbf{x}\rangle, y) - L'(\langle \mathbf{w}, \mathbf{x}\rangle, y)\,\langle \mathbf{u}, \mathbf{x}\rangle \leq \frac{1}{2}\,\beta(\langle \mathbf{u}, \mathbf{x}\rangle)^2$. Taking expectation over $(\mathbf{x}, y)$ and noting that $\nabla R(\mathbf{w}) = \mathbb{E}[L'(\langle \mathbf{w}, \mathbf{x}\rangle, y)\,\mathbf{x}]$ we get

$$R(\mathbf{w}+\mathbf{u}) - R(\mathbf{w}) - \langle \nabla R(\mathbf{w}), \mathbf{u}\rangle \leq \frac{\beta}{2}\mathbb{E}[\langle \mathbf{u}, \mathbf{x}\rangle^2] \ .$$

This gives the first inequality in the lemma. For the second inequality we use Hölder inequality and the assumption $\|\mathbf{x}\|_\infty \leq 1$ to get that $\mathbb{E}[\langle \mathbf{u}, \mathbf{x}\rangle^2] \leq \mathbb{E}[\|\mathbf{u}\|_1^2\,\|\mathbf{x}\|_\infty^2] \leq \|\mathbf{u}\|_1^2$. □

LEMMA B.2. *Let $r > 0$ and let $\epsilon_0, \epsilon_1, \ldots$ be a sequence such that $\epsilon_{t+1} \leq \epsilon_t - r\,\epsilon_t^2$ for all t. Let ϵ be a positive scalar and k be a positive integer such that $k \geq \left\lceil \frac{1}{r\,\epsilon} \right\rceil$, then, $\epsilon_k \leq \epsilon$.*

Proof. We have

$$\epsilon_1 \leq \epsilon_0 - r\epsilon_0^2 \leq 1/(4r),$$

where the maximum is achieved at $\epsilon_0 = 1/(2r)$.

Next, we use an inductive argument to show that for $t \geq 1$ we have

$$\epsilon_t \leq \frac{1}{r(t+1)} \ , \tag{B.1}$$

which will imply the desired bound in the lemma. Equation (B.1) clearly holds for $t = 1$. Assume that it holds for some $t \geq 1$. Let $\eta_t = r\epsilon_t$, so we know that $\eta_t \leq 1/(t+1)$ and we need to show that $\eta_{t+1} \leq 1/(t+2)$. The assumption $\epsilon_{t+1} \leq \epsilon_t - r\epsilon_t^2$ gives

$$\eta_{t+1} \leq \eta_t - \eta_t^2 = \eta_t(1-\eta_t) \leq \frac{\eta_t}{1+\eta_t} = \frac{1}{1+1/\eta_t} \leq \frac{1}{t+2} \ ,$$

where the first inequality is because $1 \geq 1 - \eta_t^2 = (1-\eta_t)(1+\eta_t)$ and the second inequality follows from the inductive assumption. □

The following lemma 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 B.3. *Let $\mathbf{x} = (x_1, \ldots, x_d)$ be a sequence of independent Bernoulli random variables with $0.05 \leq \Pr[x_k = 1] \leq 0.95$. Let Q be an arbitrary polynomial over d variables of degree r. Then,*

$$\mathbb{E}[\,|Q(\mathbf{x})|\,] \ \geq \ (0.2)^r \ \mathbb{E}[\,|Q(\mathbf{x})|^2\,]^{\frac{1}{2}} \quad .$$

Proof.

Using Hölder's inequality with $p = 3/2$ and $q = 3$ we have

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

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

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

We now use Corollary (3.2) from (21) to get that

$$\mathbb{E}[|Q(\mathbf{x})|^2]^{\frac{1}{2}} \geq \sigma_{4,2}(\alpha)^r \mathbb{E}[|Q(\mathbf{x})|^4]^{\frac{1}{4}} \ ,$$

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 Equation (B.2) and noting that for $\alpha \in (.05, .5)$ we have $\sigma_{4,2}(\alpha)^2 \geq 0.2$. □