Generalization of ERM in Stochastic Convex Optimization: The Dimension Strikes Back

Vitaly Feldman

IBM Research – Almaden

Abstract

In stochastic convex optimization the goal is to minimize a convex function $F(x) \doteq \mathbf{E}_{\mathbf{f} \sim D}[\mathbf{f}(x)]$ over a convex set $\mathcal{K} \subset \mathbb{R}^d$ where D is some unknown distribution and each $f(\cdot)$ in the support of D is convex over $\mathcal{K}$. The optimization is commonly based on i.i.d. samples $f^1, f^2, \ldots, f^n$ from D. A standard approach to such problems is empirical risk minimization (ERM) that optimizes $F_S(x) \doteq \frac{1}{n} \sum_{i \leq n} f^i(x)$. Here we consider the question of how many samples are necessary for ERM to succeed and the closely related question of uniform convergence of F_S to F over $\mathcal{K}$. We demonstrate that in the standard ℓ_p/ℓ_q setting of Lipschitz-bounded functions over a $\mathcal{K}$ of bounded radius, ERM requires sample size that scales linearly with the dimension d. This nearly matches standard upper bounds and improves on $\Omega(\log d)$ dependence proved for ℓ_2/ℓ_2 setting in [SSSS09]. In stark contrast, these problems can be solved using dimension-independent number of samples for ℓ_2/ℓ_2 setting and $\log d$ dependence for ℓ_1/ℓ_∞ setting using other approaches.

We further show that our lower bound applies even if the functions in the support of D are smooth and efficiently computable and even if an ℓ_1 regularization term is added. Finally, we demonstrate that for a more general class of bounded-range (but not Lipschitz-bounded) stochastic convex programs an infinite gap appears already in dimension 2.

1   Introduction

Numerous central problems in machine learning, statistics and operations research are special cases of stochastic optimization from i.i.d. data samples. In this problem the goal is to optimize the value of the expected objective function $F(x) \doteq \mathbf{E}_{\mathbf{f} \sim D}[\mathbf{f}(x)]$ over some set $\mathcal{K}$ given i.i.d. samples $f^1, f^2, \ldots, f^n$ of $\mathbf{f}$. For example, in supervised learning the set $\mathcal{K}$ consists of hypothesis functions from Z to Y and each sample is an example described by a pair $(z, y) \in (Z, Y)$. For some fixed loss function $L : Y \times Y \to \mathbb{R}$, an example (z, y) defines a function from $\mathcal{K}$ to $\mathbb{R}$ given by $f_{(z,y)}(h) = L(h(z), y)$. The goal is to find a hypothesis h that (approximately) minimizes the expected loss relative to some distribution P over examples: $\mathbf{E}_{(z,y) \sim P}[L(h(z), y)] = \mathbf{E}_{(z,y) \sim P}[f_{(z,y)}(h)]$.

Here we are interested in stochastic convex optimization (SCO) problems in which $\mathcal{K}$ is some convex subset of $\mathbb{R}^d$ and each function in the support of D is convex over $\mathcal{K}$. The importance of this setting stems from the fact that such problems can be solved efficiently via a large variety of known techniques. Therefore in many applications even if the original optimization problem is not convex, it is replaced by a convex relaxation.

A classic and widely-used approach to solving stochastic optimization problems is empirical risk minimization (ERM) also referred to as stochastic average approximation (SAA) in the optimization literature. In this approach, given a set of samples $S = (f^1, f^2, \ldots, f^n)$ the empirical objective function:

$F_S(x) \doteq \frac{1}{n} \sum_{i \leq n} f^i(x)$ is optimized (sometimes with an additional regularization term such as $\lambda \|x\|^2$ for some $\lambda > 0$). The question we address here is the number of samples required for this approach to work *distribution-independently*. More specifically, for some fixed convex body $\mathcal{K}$ and fixed set of convex functions $\mathcal{F}$ over $\mathcal{K}$, what is the smallest number of samples n such that for every probability distribution D supported on $\mathcal{F}$, any algorithm that minimizes F_S given n i.i.d. samples from D will produce an ϵ-optimal solution $\hat{x}$ to the problem (namely, $F(\hat{x}) \leq \min_{x \in \mathcal{K}} F(x) + \epsilon$) with probability at least $1 - \delta$? We will refer to this number as the sample complexity of ERM for ϵ-optimizing $\mathcal{F}$ over $\mathcal{K}$ (we will fix $\delta = 1/2$ for now).

The sample complexity of ERM for ϵ-optimizing $\mathcal{F}$ over $\mathcal{K}$ is lower bounded by the sample complexity of ϵ-optimizing $\mathcal{F}$ over $\mathcal{K}$, that is the number of samples that is necessary to find an ϵ-optimal solution for any algorithm. On the other hand, it is upper bounded by the number of samples that ensures uniform convergence of F_S to F. Namely, if with probability $\geq 1 - \delta$, for all $x \in \mathcal{K}$, $|F_S(x) - F(x)| \leq \epsilon/2$ then, clearly, any algorithm based on ERM will succeed. As a result, ERM and uniform convergence are the primary tool for analysis of the sample complexity of learning problems and are the key subject of study in statistical learning theory. Fundamental results in VC theory imply that in some settings, such as binary classification and least-squares regression, uniform convergence is also a necessary condition for learnability (*e.g.* [Vap98, SSBD14]) and therefore the three measures of sample complexity mentioned above nearly coincide.

In the context of stochastic convex optimization the study of sample complexity of ERM and uniform convergence was initiated in a groundbreaking work of Shalev-Shwartz, Shamir, Srebro and Sridharan [SSSS09]. They demonstrated that the relationships between these notions of sample complexity are substantially more delicate even in the most well-studied settings of SCO. Specifically, let $\mathcal{K}$ be a unit ℓ_2 ball and $\mathcal{F}$ be the set of all convex sub-differentiable functions with Lipschitz constant relative to ℓ_2 bounded by 1 or, equivalently, $\|\nabla f(x)\|_2 \leq 1$ for all $x \in \mathcal{K}$. Then, known algorithm for SCO imply that sample complexity of this problem is $O(1/\epsilon^2)$ and often expressed as $1/\sqrt{n}$ rate of convergence (*e.g.* [NJLS09, SSBD14]). On the other hand, Shalev-Shwartz *et al.*[SSSS09] show[1] that the sample complexity of ERM for solving this problem with $\epsilon = 1/2$ is $\Omega(\log d)$. The only known upper bound for sample complexity of ERM is $\tilde{O}(d/\epsilon^2)$ and relies only on the uniform convergence of Lipschitz-bounded functions [SN05, SSSS09].

As can seen from this discussion, the work of Shalev-Shwartz *et al.*[SSSS09] still leaves a major gap between known bounds on sample complexity of ERM (and also uniform convergence) for this basic Lipschitz-bounded ℓ_2/ℓ_2 setup. Another natural question is whether the gap is present in the popular ℓ_1/ℓ_∞ setup. In this setup $\mathcal{K}$ is a unit ℓ_1 ball (or in some cases a simplex) and $\|\nabla f(x)\|_\infty \leq 1$ for all $x \in \mathcal{K}$. The sample complexity of SCO in this setup is $\theta(\log d/\epsilon^2)$ (*e.g.* [NJLS09, SSBD14]) and therefore, even an appropriately modified lower bound in [SSSS09], does not imply any gap. More generally, the choice of norm can have a major impact on the relationship between these sample complexities and hence needs to be treated carefully. For example, for (the reversed) ℓ_∞/ℓ_1 setting the sample complexity of the problem is $\theta(d/\epsilon^2)$ (*e.g.* [FGV15]) and nearly coincides with the number of samples sufficient for uniform convergence.

1.1 Overview of Results

In this work we substantially strengthen the lower bound in [SSSS09] proving that a linear dependence on the dimension d is necessary for ERM (and, consequently, uniform convergence). We then extend the lower bound to all ℓ_p/ℓ_q setups and examine several related questions. Finally, we examine a more general setting of bounded-range SCO (that is $|f(x)| \leq 1$ for all $x \in \mathcal{K}$). While the sample complexity of this setting is still low (for example $\tilde{O}(1/\epsilon^2)$ when $\mathcal{K}$ is an ℓ_2 ball) and efficient algorithms are known, we show that ERM

---

[1]The dependence on d is not stated explicitly but follows immediately from their analysis.

might require an infinite number of samples already for $d = 2$.

Our work implies that in SCO, even optimization algorithms that exactly minimize the empirical objective function can produce solutions with generalization error that is much larger than the generalization error of solutions obtained via some standard approaches. Another, somewhat counterintuitive, conclusion from our lower bounds is that, from the point of view of generalization of ERM and uniform convergence, convexity does not reduce the sample complexity in the worst case.

Basic construction: Our basic construction is fairly simple and its analysis is inspired by the technique in [SSSS09]. It is based on functions of the form $\max\{1/2, \max_{v \in V} \langle v, x \rangle\}$. Note that the maximum operator preserves both convexity and Lipschitz bound (relative to any norm). See Figure 1 for an illustration of such function for $d = 2$.
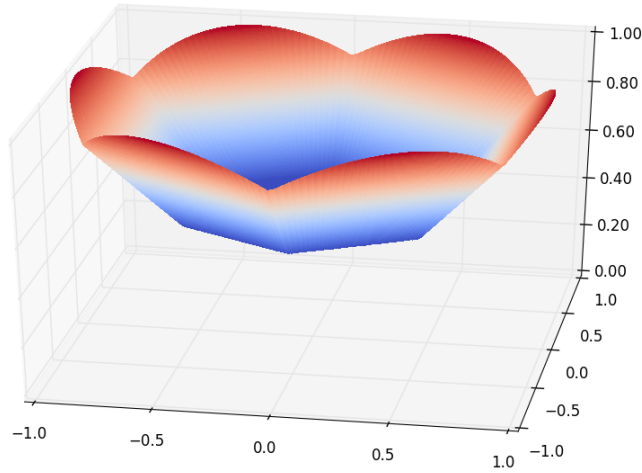

Figure 1: Basic construction for $d = 2$.

The distribution over the sets V that define such functions is uniform over all subsets of some set of vectors W of size $2^{d/6}$ such that for any two district $u, v \in W$, $\langle u, v \rangle \leq 1/2$. Equivalently, each element of W is included in V with probability $1/2$ independently of other elements in W. This implies that if the number of samples is less than $d/6$ then, with probability $> 1/2$, at least one of the vectors in W (say w) will not be observed in any of the samples. This implies that F_S can be minimized while maximizing $\langle w, x \rangle$ (the maximum over the unit ℓ_2 ball is w). Note that a function randomly chosen from our distribution includes the term $\langle w, x \rangle$ in the maximum operator with probability $1/2$. Therefore the value of the expected function F at w is $3/4$ whereas the minimum of F is $1/2$. In particular, there exists an ERM algorithm with generalization error of at least $1/4$. The details of the construction appear in Sec. 3.1 and Thm. 3.3 gives the formal statement of the lower bound. We also show that, by scaling the construction appropriately, we can obtain the same lower bound for any ℓ_p/ℓ_q setup with $1/p + 1/q = 1$ (see Thm. 3.5).

Low complexity construction: The basic construction relies on functions that require $2^{d/6}$ bits to describe and exponential time to compute. Most application of SCO use efficiently computable functions and therefore it is natural to ask whether the lower bound still holds for such functions. To answer this question we describe a construction based on a set of functions where each function requires just $\log d$ bits to describe (there are at most $d/2$ functions in the support of the distribution) and each function can be computed in $O(d)$ time. To achieve this we will use W that consists of (scaled) codewords of an asymptotically good and

efficiently computable binary error-correcting code [Jus72, Spi96]. The functions are defined in a similar way but the additional structure of the code allows to use at most $d/2$ subsets of W to define the functions. Further details of the construction appear in Section 4.

Smoothness: The use of maximum operator results in functions that are highly non-smooth (that is, their gradient is not Lipschitz-bounded) whereas the construction in [SSSS09] uses smooth functions. Smoothness plays a crucial role in many algorithms for convex optimization (see [Bub15] for examples). It reduces the sample complexity of SCO in ℓ_2/ℓ_2 setup to $O(1/\epsilon)$ when the smoothness parameter is a constant (*e.g.* [NJLS09, SSBD14]). Therefore it is natural to ask whether our strong lower bound holds for smooth functions as well. We describe a modification of our construction that proves a similar lower bound in the smooth case (with generalization error of $1/128$). The main idea is to replace each linear function $\langle v, x \rangle$ with some smooth function $\nu(\langle v, x \rangle)$ guaranteeing that for different vectors $v^1, v^2 \in W$ and every $x \in \mathcal{K}$, only one of $\nu(\langle v^1, x \rangle)$ and $\nu(\langle v^2, x \rangle)$ can be non-zero. This allows to easily control the smoothness of $\max_{v \in V} \nu(\langle v, x \rangle)$. See Figure 2 for an illustration of a function on which the construction is based (for $d = 2$). The details of this construction appear in Sec. 3.2 and the formal statement in Thm. 3.7.
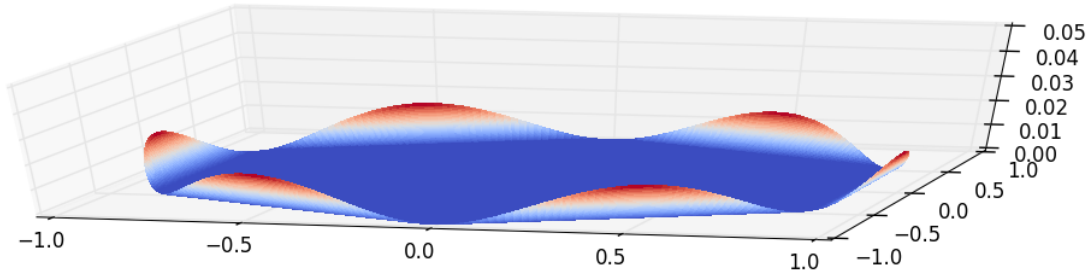

Figure 2: Construction using 1-smooth functions for $d = 2$.

ℓ_1**-regularization:** Another important contribution in [SSSS09] is the demonstration of the important role that strong convexity plays for generalization in SCO: Minimization of $F_S(x) + \lambda R(x)$ ensures that ERM will have low generalization error whenever $R(x)$ is strongly convex (for a sufficiently large λ). This result is based on the proof that ERM of a strongly convex Lipschitz function is *uniform replace-one* stable and the connection between such stability and generalization showed in [BE02] (see also [SSSSS10] for a detailed treatment of the relationship between generalization and stability). It is natural to ask whether other approaches to regularization will ensure generalization. We demonstrate that for the commonly used ℓ_1 regularization the answer is negative. We prove this using a simple modification of our lower bound construction: We shift the functions to the positive orthant where the regularization terms $\lambda \|x\|_1$ is just a linear function. We then subtract this linear function from each function in our construction, thereby balancing the regularization (while maintaining convexity and Lipschitz-boundedness). The details of this construction appear in Sec. 3.3 (see Thm. 3.8).

Dependence on accuracy: For simplicity and convenience we have ignored the dependence on the accuracy ϵ, Lipschitz bound L and radius R of $\mathcal{K}$ in our lower bounds. It is easy to see, that this more general setting can be reduced to the case we consider here (Lipschitz bound and radius are equal to 1) with accuracy parameter $\epsilon' = \epsilon/(LR)$. We generalize our lower bound to this setting and prove that $\Omega(d/\epsilon'^2)$ samples are necessary for uniform convergence and $\Omega(d/\epsilon')$ samples are necessary for generalization of ERM. Note

that the upper bound on the sample complexity of these settings is $\tilde{O}(d/\epsilon'^2)$ and therefore the dependence on ϵ' in our lower bound does not match the upper bound for ERM. Resolving this gap or even proving any $\omega(d/\epsilon + 1/\epsilon'^2)$ lower bound is an interesting open problem. Additional details can be found in Section 3.4.

Bounded-range SCO: Finally, we consider a more general class of bounded-range convex functions Note that the Lipschitz bound of 1 and the bound of 1 on the radius of $\mathcal{K}$ imply a bound of 1 on the range (up to a constant shift which does not affect the optimization problem). While this setting is not as well-studied, efficient algorithms for it are known. For example, the online algorithm in a recent work of Rakhlin and Sridharan [RS15] together with standard online-to-batch conversion arguments [CCG04], imply that the sample complexity of this problem is $\tilde{O}(1/\epsilon^2)$ for any $\mathcal{K}$ that is an ℓ_2 ball (of any radius). For general convex bodies $\mathcal{K}$, the problems can be solved via random walk-based approaches [BLNR15, FGV15] or an adaptation of the center-of-gravity method given in [FGV15]. Here we show that for this setting ERM might completely fail already for $\mathcal{K}$ being the unit 2-dimensional ball. The construction is based on ideas similar to those we used in the smooth case and is formally described in Sec. 5. See Figure 3 for an illustration of a function used in this construction.
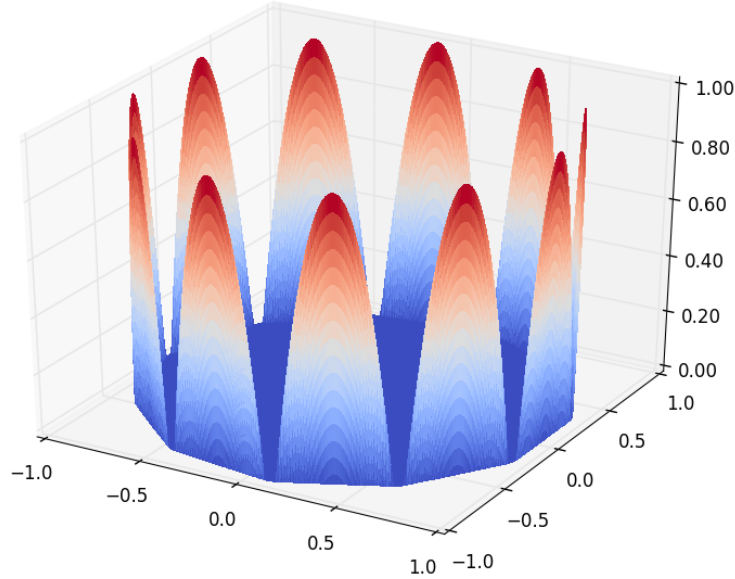

Figure 3: Construction using non-Lipschitz convex functions with range in $[0, 1]$.

2 Preliminaries

For an integer $n \geq 1$ let $[n] \doteq \{1, \ldots, n\}$. Random variables are denoted by bold letters, e.g., $\mathbf{f}$. Given $p \in [1, \infty]$ we denote the ball of radius $R > 0$ in ℓ_p norm by $\mathcal{B}_p^d(R)$, and the unit ball by $\mathcal{B}_p^d$.

For a convex body (i.e., compact convex set with nonempty interior) $\mathcal{K} \subseteq \mathbb{R}^d$, we consider problems of the form

$$\min_{\mathcal{K}}(F_D) \doteq \min_{x \in \mathcal{K}} \left\{ F_D(x) \doteq \mathop{\mathbf{E}}_{\mathbf{f} \sim D}[\mathbf{f}(x)] \right\},$$

where $\mathbf{f}$ is a random variable defined over some set of convex, sub-differentiable functions $\mathcal{F}$ on $\mathcal{K}$ and distributed according to some unknown probability distribution D. We denote $F^* = \min_{\mathcal{K}}(F_D)$. For an approximation parameter $\epsilon > 0$ the goal is to find $x \in \mathcal{K}$ such that $F_D(x) \leq F^* + \epsilon$ and we call any such x an ϵ-*optimal solution*. For an n-tuple of functions $S = (f^1, \ldots, f^n)$ we denote by $F_S \doteq \frac{1}{n} \sum_{i \in [n]} f^i$.

We say that a point $\hat{x}$ is an empirical risk minimum for an n-tuple S of functions over $\mathcal{K}$, if $F_S(\hat{x}) = \min_{\mathcal{K}}(F_S)$. In some cases there are many points that minimize F_S and in this case we refer to a specific algorithm that selects one of the minimums of F_S as an empirical risk minimizer. To make this explicit we refer to the output of such a minimizer by $\hat{x}(S)$.

Given $x \in \mathcal{K}$, and a convex function f we denote by $\nabla f(x) \in \partial f(x)$ an arbitrary selection of a subgradient. Let us make a brief reminder of some important classes of convex functions. Let $p \in [1, \infty]$ and $q = p_* \doteq 1/(1 - 1/p)$. We say that a subdifferentiable convex function $f : \mathcal{K} \to \mathbb{R}$ is in the class

- $\mathcal{F}(\mathcal{K}, B)$ of B-bounded-range functions if for all $x \in \mathcal{K}$, $|f(x)| \leq B$.

- $\mathcal{F}_p^0(\mathcal{K}, L)$ of L-Lipschitz continuous functions w.r.t. ℓ_p, if for all $x, y \in \mathcal{K}$, $|f(x) - f(y)| \leq L\|x - y\|_p$;

- $\mathcal{F}_p^1(\mathcal{K}, \sigma)$ of functions with σ-Lipschitz continuous gradient w.r.t. ℓ_p, if for all $x, y \in \mathcal{K}$, $\|\nabla f(x) - \nabla f(y)\|_q \leq \sigma\|x - y\|_p$.

We will omit p from the notation when $p = 2$. .

3   Lower Bounds for Lipschitz-Bounded SCO

In this section we present our main lower bounds for SCO of Lipschitz-bounded convex functions. For comparison purposes we start by formally stating some known bounds on sample complexity of solving such problems. The following uniform convergence bounds can be easily derived from the standard covering number argument (*e.g.* [SN05, SSSS09])

Theorem 3.1. *For $p \in [1, \infty]$, let $\mathcal{K} \subseteq \mathcal{B}_p^d(R)$ and let D be any distribution supported on functions L-Lipschitz on $\mathcal{K}$ relative to ℓ_p (not necessarily convex). Then, for every $\epsilon, \delta > 0$ and $n \geq n_1 = O\left(\frac{d \cdot (LR)^2 \cdot \log(dLR/(\epsilon\delta))}{\epsilon^2}\right)$*

$$\Pr_{\mathbf{S} \sim D^n} [\exists x \in \mathcal{K}, \ |F_D(x) - F_{\mathbf{S}}(x)| \geq \epsilon] \leq \delta.$$

The following upper bounds on sample complexity of Lipschitz-bounded SCO can be obtained from several known algorithms [NJLS09, SSSS09] (see [SSBD14] for a textbook exposition for $p = 2$).

Theorem 3.2. *For $p \in [1, 2]$, let $\mathcal{K} \subseteq \mathcal{B}_p^d(R)$. Then, there is an algorithm $\mathcal{A}_p$ that given $\epsilon, \delta > 0$ and $n = n_p(d, R, L, \epsilon, \delta)$ i.i.d. samples from any distribution D supported on $\mathcal{F}_p^0(\mathcal{K}, L)$, outputs an ϵ-optimal solution to F_D over $\mathcal{K}$ with probability $\geq 1 - \delta$. For $p \in (1, 2]$, $n_p = O((LR/\epsilon)^2 \cdot \log(1/\delta))$ and for $p = 1$, $n_p = O((LR/\epsilon)^2 \cdot \log d \cdot \log(1/\delta))$.*

Stronger results are known under additional assumptions on smoothness and/or strong convexity (*e.g.* [NJLS09, RSS12, SZ13, BM13]).

3.1 Non-smooth construction

We will start with a simpler lower bound for non-smooth functions. For simplicity, we will also restrict $R = L = 1$. Lower bounds for the general setting can be easily obtained from this case by scaling the domain and desired accuracy(see Thm. 3.10 for additional details).

We will need a set of vectors $W \subseteq \{-1,1\}^d$ with the following property: for any distinct $w^1, w^2 \in W$, $\langle w^1, w^2 \rangle \leq d/2$. The Chernoff bound together with a standard packing argument imply that there exists a set W with this property of size $\geq e^{d/8} \geq 2^{d/6}$.

For any subset V of W we define a function

$$g_V(x) \doteq \max\{1/2, \max_{w \in V} \langle \bar{w}, x \rangle\}, \tag{1}$$

where $\bar{w} \doteq w/\|w\| = w/\sqrt{d}$. See Figure 1 for an illustration. We first observe that g_V is convex and 1-Lipschitz (relative to ℓ_2). This immediately follows from $\langle \bar{w}, x \rangle$ being convex and 1-Lipschitz for every w and g_V being the maximum of convex and 1-Lipschitz functions.

Theorem 3.3. *Let $\mathcal{K} = \mathcal{B}_2^d$ and we define $\mathcal{H}_2 \doteq \{g_V \mid V \subseteq W\}$ for g_V defined in eq. (1). Let D be the uniform distribution over $\mathcal{H}_2$. Then for $n \leq d/6$ and every set of samples S there exists an ERM $\hat{x}(S)$ such that*

$$\Pr_{S \sim D^n} [F_D(\hat{x}(\mathbf{S})) - F^* \geq 1/4] > 1/2.$$

Proof. We start by observing that the uniform distribution over $\mathcal{H}_2$ is equivalent to picking the function $g_{\mathbf{V}}$ where $\mathbf{V}$ is obtained by including every element of W with probability $1/2$ randomly and independently of all other elements. Further, by the properties of W, for every $w \in W$, and $V \subseteq W$, $g_V(\bar{w}) = 1$ if $w \in V$ and $g_V(\bar{w}) = 1/2$ otherwise. For $g_{\mathbf{V}}$ chosen randomly with respect to D, we have that $w \in \mathbf{V}$ with probability exactly $1/2$. This implies that $F_D(\bar{w}) = 3/4$.

Let $\mathbf{S} = (g_{\mathbf{V}_1}, \ldots, g_{\mathbf{V}_n})$ be the random samples. Observe that $\min_{\mathcal{K}}(F_\mathbf{S}) = 1/2$ and $F^* = \min_{\mathcal{K}}(F_D) = 1/2$ (the minimum is achieved at the origin $\bar{0}$). Now, if $\bigcup_{i \in [n]} \mathbf{V}_i \neq W$ then let $\hat{x}(\mathbf{S}) \doteq \bar{w}$ for any $w \in W \setminus \bigcup_{i \in [n]} \mathbf{V}_i$. Otherwise $\hat{x}(\mathbf{S})$ is defined to be the origin $\bar{0}$. Then by the property of $\mathcal{H}_2$ mentioned above, we have that for all i, $g_{\mathbf{V}_i}(\hat{x}(\mathbf{S})) = 1/2$ and hence $F_\mathbf{S}(\hat{x}(\mathbf{S})) = 1/2$. This means that $\hat{x}(\mathbf{S})$ is a minimizer of $F_\mathbf{S}$.

Combining these statements, we get that, if $\bigcup_{i \in [n]} \mathbf{V}_i \neq W$ then there exists an ERM $\hat{x}(\mathbf{S})$ such that $F_\mathbf{S}(\hat{x}(\mathbf{S})) = \min_{\mathcal{K}}(F_\mathbf{S})$ and $F_D(\hat{x}(\mathbf{S})) - F^* = 1/4$. Therefore to prove the claim it suffices to show that for $n \leq d/6$ we have that

$$\Pr_{S \sim D^n} \left[\bigcup_{i \in [n]} \mathbf{V}_i \neq W \right] > \frac{1}{2}.$$

This easily follows from observing that for the uniform distribution over subsets of W, for every $w \in W$,

$$\Pr_{S \sim D^n} \left[w \in \bigcup_{i \in [n]} \mathbf{V}_i \right] = 1 - 2^{-n}$$

and this event is independent from the inclusion of other elements in $\bigcup_{i \in [n]} \mathbf{V}_i$. Therefore

$$\Pr_{S \sim D^n} \left[\bigcup_{i \in [n]} \mathbf{V}_i = W \right] = \left(1 - 2^{-n}\right)^{|W|} \leq e^{-2^{-n} \cdot 2^{d/6}} \leq e^{-1} < \frac{1}{2}.$$

7

$\square$

Remark 3.4. *In our construction there is a different ERM algorithm that does solve the problem (and generalizes well). For example, the algorithm that always outputs the origin $\bar{0}$. Therefore it is natural to ask whether the same lower bound holds when there exists a unique minimizer. Shalev-Shwartz et al.[SSSS09] show that their lower bound construction can be slightly modified to ensure that the minimizer is unique while still having large generalization error. An analogous modification appears to be much harder to analyze in our construction and it is unclear to us how to ensure uniqueness in our strong lower bounds. A further question in this direction is whether it is possible to construct a distribution for which the empirical minimizer with large generalization error is unique and its value is noticeably (at least by $1/\mathrm{poly}(d)$) smaller than the value of F_S at any point x that generalizes well. Such distribution would imply that the solutions that "overfits" can be found easily (for example, in a polynomial number of iterations of the gradient descent).*

Other ℓ_p norms: We now observe that exactly the same approach can be used to extend this lower bound to ℓ_p/ℓ_q setting. Specifically, for $p \in [1, \infty]$ and $q = p_*$ we define

$$g_{p,V}(x) \doteq \max \left\{ \frac{1}{2}, \max_{w \in V} \frac{\langle w, x \rangle}{d^{1/q}} \right\}.$$

It is easy to see that for every $V \subseteq W$, $g_{q,V} \in \mathcal{F}_p^0(\mathcal{B}_p^d, 1)$. We can now use the same argument as before with the appropriate normalization factor for points in $\mathcal{B}_p^d$. Namely, instead of $\bar{w}$ for $w \in W$ we consider the values of the minimized functions at $w/d^{1/p} \in \mathcal{B}_p^d$. This gives the following generalization of Thm. 3.3.

Theorem 3.5. *For every $p \in [1, \infty]$ let $\mathcal{K} = \mathcal{B}_p^d$ and we define $\mathcal{H}_p \doteq \{g_{p,V} \mid V \subseteq W\}$ and let D be the uniform distribution over $\mathcal{H}_p$. Then for $n \le d/6$ and every set of samples S there exists an ERM $\hat{x}(S)$ such that*

$$\Pr_{\mathbf{S} \sim D^n} [F_D(\hat{x}(\mathbf{S})) - F^* \ge 1/4] > 1/2.$$

3.2 Smoothness does not help

We now extend the lower bound to smooth functions. We will for simplicity restrict our attention to ℓ_2 but analogous modifications can be made for other ℓ_p norms. The functions g_V that we used in the construction use two maximum operators each of which introduces non-smoothness. To deal with maximum with $1/2$ we simply replace the function $\max\{1/2, \langle \bar{w}, x \rangle\}$ with a quadratically smoothed version (in the same way as hinge loss is sometimes replaced with modified Huber loss). To deal with the maximum over all $w \in V$, we show that it is possible to ensure that individual components do not "interact". That is, at every point x, the value, gradient and Hessian of at most one component function are non-zero (value, vector and matrix, respectively). This ensures that maximum becomes addition and Lipschitz/smoothness constants can be upper-bounded easily.

Formally, we define

$$\nu(a) \doteq \begin{cases} 0 & \text{if } a \le 0 \\ a^2 & \text{otherwise.} \end{cases}$$

Now, for $V \subseteq W$, we define

$$h_V(x) \doteq \sum_{w \in V} \nu(\langle \bar{w}, x \rangle - 7/8). \tag{2}$$

See Figure 2 for an illustration. We first prove that h_V is $1/4$-Lipschitz and 1-smooth.

Lemma 3.6. *For every $V \subseteq W$ and h_V defined in eq. (2) we have $h_V \in \mathcal{F}_2^0(\mathcal{B}_2^d, 1/4) \cap \mathcal{F}_2^1(\mathcal{B}_2^d, 1)$.*

Proof. It is easy to see that $\nu(\langle \bar{w}, x \rangle - 7/8)$ is convex for every w and hence h_V is convex. Next we observe that for every point $x \in \mathcal{B}_2^d$, there is at most one $w \in W$ such that $\langle \bar{w}, x \rangle > 7/8$. If $\langle \bar{w}, x \rangle > 7/8$ then $\|\bar{w} - x\|^2 = \|\bar{w}\|^2 + \|x\|^2 - 2\langle \bar{w}, x \rangle < 1 + 1 - 2(7/8) = 1/4$. On the other hand, by the properties of W, for distinct w^1, w^2 we have that $\|\bar{w}^1 - \bar{w}^2\|^2 = 2 - 2\langle \bar{w}^1, \bar{w}^2 \rangle \geq 1$. Combining these bounds on distances we obtain that if we assume that $\langle \bar{w}^1, x \rangle > 7/8$ and $\langle \bar{w}^2, x \rangle > 7/8$ then we obtain a contradiction

$$\|\bar{w}^1 - \bar{w}^2\| \leq \|\bar{w}^1 - x\| + \|\bar{w}^2 - x\| < 1.$$

From here we can conclude that

$$\nabla h_V(x) = \begin{cases} 2(\langle \bar{w}, x \rangle - 7/8) \cdot \bar{w} & \text{if } \exists w \in V, \ \langle \bar{w}, x \rangle > 7/8 \\ 0 & \text{otherwise} \end{cases}.$$

This immediately implies that $\|\nabla h_V(x)\| \leq 1/4$ and hence h_V is $1/4$-Lipschitz.

We now prove smoothness. Given two points $x, y \in \mathcal{B}_2^d$ we consider two cases. First the simpler case when there is at most one $w \in V$ such that either $\langle \bar{w}, x \rangle > 7/8$ or $\langle \bar{w}, y \rangle > 7/8$. In this case $\nabla h_V(x) = \nabla \nu(\langle \bar{w}, x \rangle - 7/8)$ and $\nabla h_V(y) = \nabla \nu(\langle \bar{w}, y \rangle - 7/8)$. This implies that the 1-smoothness condition is implied by 1-smoothness of $\nu(\langle \bar{w}, \cdot \rangle - 7/8)$. That is one can easily verify that $\|\nabla h_V(x) - \nabla h_V(y)\| \leq \|x - y\|$.

Next we consider the case where for x there is $w^1 \in V$ such that $\langle \bar{w}^1, x \rangle > 7/8$, for y there is $w^2 \in V$ such that $\langle \bar{w}^2, y \rangle > 7/8$ and $w^1 \neq w^2$. Then there exists a point $z \in \mathcal{B}_2^d$ on the line connecting x and y such that $\langle \bar{w}^1, z \rangle \leq 7/8$ and $\langle \bar{w}^2, z \rangle \leq 7/8$. Clearly, $\|x - y\| = \|x - z\| + \|z - y\|$. On the other hand, by the analysis of the previous case we have that $\|\nabla h_V(x) - \nabla h_V(z)\| \leq \|x - z\|$ and $\|\nabla h_V(z) - \nabla h_V(y)\| \leq \|z - y\|$. Combining these inequalities we obtain that

$$\|\nabla h_V(x) - \nabla h_V(y)\| \leq \|\nabla h_V(x) - \nabla h_V(z)\| + \|\nabla h_V(z) - \nabla h_V(y)\| \leq \|x - z\| + \|z - y\| = \|x - y\|.$$

$\square$

From here we can use the proof approach from Thm. 3.3 but with h_V in place of g_V.

Theorem 3.7. *Let $\mathcal{K} = \mathcal{B}_2^d$ and we define $\mathcal{H} \doteq \{h_V \mid V \subseteq W\}$ for h_V defined in eq. (2). Let D be the uniform distribution over $\mathcal{H}$. Then for $n \leq d/6$ and every set of samples S there exists an ERM $\hat{x}(S)$ such that*

$$\Pr_{\mathbf{S} \sim D^n} [F_D(\hat{x}(\mathbf{S})) - F^* \geq 1/128] > 1/2.$$

Proof. Let $\mathbf{S} = (h_{\mathbf{V}_1}, \ldots, h_{\mathbf{V}_n})$ be the random samples. As before we first note that $\min_{\mathcal{K}}(F_{\mathbf{S}}) = 0$ and $F^* = 0$. Further, for every $w \in W$, $h_V(\bar{w}) = 1/64$ if $w \in V$ and $h_V(\bar{w}) = 0$ otherwise. Hence $F_D(\bar{w}) = 1/128$. Now, if $\bigcup_{i \in [n]} \mathbf{V}_i \neq W$ then let $\hat{x}(\mathbf{S}) \doteq \bar{w}$ for some $w \in W \setminus \bigcup_{i \in [n]} \mathbf{V}_i$. Then for all i, $h_{\mathbf{V}_i}(\hat{x}(\mathbf{S})) = 0$ and hence $F_{\mathbf{S}}(\hat{x}(\mathbf{S})) = 0$. This means that $\hat{x}(\mathbf{S})$ is a minimizer of $F_{\mathbf{S}}$ and $F_D(\hat{x}(\mathbf{S})) - F^* = 1/128$.

Now, exactly as in Thm. 3.3, we can conclude that $\bigcup_{i \in [n]} \mathbf{V}_i \neq W$ with probability $> 1/2$. $\square$

9

3.3 ℓ_1 Regularization does not help

Next we show that the lower bound holds even with an additional ℓ_1 regularization term $\lambda\|x\|$ for positive $\lambda \leq 1/\sqrt{d}$. (Note that if $\lambda > 1/\sqrt{d}$ then the resulting program is no longer 1-Lipschitz relative to ℓ_2. Any constant λ can be allowed for ℓ_1/ℓ_∞ setup). To achieve this we shift the construction to the positive orthant (that is x such that $x_i \geq 0$ for all $i \in [d]$). In this orthant the subgradient of the regularization term is simply $\lambda\bar{1}$ where $\bar{1}$ is the all 1's vector. We can add a linear term to each function in our distribution that balances this term thereby reducing the analysis to non-regularized case. More formally, we define the following family of functions. For $V \subseteq W$,

$$h_V^\lambda(x) \doteq h_V(x - \bar{1}/\sqrt{d}) - \lambda\langle\bar{1}, x\rangle. \tag{3}$$

Note that over $\mathcal{B}_2^d(2)$, $h_V^\lambda(x)$ is L-Lipschitz for $L \leq 2(2 - 7/8) + \lambda\sqrt{d} \leq 9/4$. We now state and prove this formally.

Theorem 3.8. *Let $\mathcal{K} = \mathcal{B}_2^d(2)$ and for a given $\lambda \in (0, 1/\sqrt{d}]$, we define $\mathcal{H}^\lambda \doteq \{h_V^\lambda \mid V \subseteq W\}$ for h_V^λ defined in eq. (3). Let D be the uniform distribution over $\mathcal{H}^\lambda$. Then for $n \leq d/6$ and every set of samples S there exists $\hat{x}(S)$ such that*

- $F_S(\hat{x}(S)) = \min_{x \in \mathcal{K}}(F_S(x) + \lambda\|x\|_1)$;

- $\mathbf{Pr}_{\mathbf{S} \sim D^n}[F_D(\hat{x}(\mathbf{S})) - F^* \geq 1/128] > 1/2$.

Proof. Let $\mathbf{S} = (h_{\mathbf{V}_1}^\lambda, \ldots, h_{\mathbf{V}_n}^\lambda)$ be the random samples. We first note that $F^* = F_D(\bar{0}) = 0$ and

$$\min_{x \in \mathcal{K}}(F_{\mathbf{S}}(x) + \lambda\|x\|_1) = \min_{x \in \mathcal{K}}\left(\sum_{i \in [n]} h_{\mathbf{V}_i}\left(x - \frac{\bar{1}}{\sqrt{d}}\right) - \lambda\langle\bar{1}, x\rangle + \lambda\|x\|_1\right)$$

$$\geq \min_{x \in \mathcal{K}}\left(\sum_{i \in [n]} h_{\mathbf{V}_i}\left(x - \frac{\bar{1}}{\sqrt{d}}\right)\right) \geq 0.$$

Further, for every $w \in W$, $\bar{w} + \bar{1}/\sqrt{d}$ is in the positive orthant and in $\mathcal{B}_2^d(2)$. Hence $h_V^\lambda(\bar{w} + \bar{1}/\sqrt{d}) = h_V(\bar{w})$. We can therefore apply the analysis from Thm. 3.7 to obtain the claim. $\qquad\square$

3.4 Dependence on ϵ

We now briefly consider the dependence of our lower bound on the desired accuracy. Note that the upper bound for uniform convergence scales as $\tilde{O}(d/\epsilon^2)$.

We first observe that our construction implies a lower bound of $\Omega(d/\epsilon^2)$ for uniform convergence nearly matching the upper bound (we do this for the simpler non-smooth ℓ_2 setting but the same applies to other setting we consider).

Theorem 3.9. *Let $\mathcal{K} = \mathcal{B}_2^d$ and we define $\mathcal{H}_2 \doteq \{g_V \mid V \subseteq W\}$ for g_V defined in eq. (1). Let D be the uniform distribution over $\mathcal{H}_2$. Then for any $\epsilon > 0$ and $n \leq n_1 = \Omega(d/\epsilon^2)$ and every set of samples S there exists a point $\hat{x}(S)$ such that*

$$\Pr_{\mathbf{S} \sim D^n}[F_D(\hat{x}(\mathbf{S})) - F_{\mathbf{S}}(\hat{x}(\mathbf{S})) \geq \epsilon] > 1/2.$$

10

Proof. For every $w \in W$,

$$F_{\mathbf{S}}(\bar{w}) = \frac{1}{n} \sum_{i \in [n]} g_{\mathbf{V}_i}(\bar{w}) = \frac{1}{2} + \frac{1}{2n} \sum_{i \in [n]} \mathbf{1}_{\{w \in \mathbf{V}_i\}},$$

where $\mathbf{1}_{\{w \in \mathbf{V}_i\}}$ is the indicator variable of w being in $\mathbf{V}_i$. If for some w, $\frac{1}{2n} \sum_{i \in [n]} \mathbf{1}_{\{w \in \mathbf{V}_i\}} \geq 1/4 + \epsilon$ then we will obtain a point $\bar{w}$ that violates the uniform convergence by ϵ. For every w, $\sum_{i \in [n]} \mathbf{1}_{\{w \in \mathbf{V}_i\}}$ is distributed according to the binomial distribution. Using a standard approximation of the partial binomial sum up to $(1/2 - 2\epsilon)n$, we obtain that for some constant $c > 0$, the probability that this sum is $\geq 1/2 + 2\epsilon$ is at least

$$\frac{1}{2^n} \cdot \frac{1}{\sqrt{8n(1/4 - \epsilon^2)}} \cdot \left(\frac{1}{2} + 2\epsilon\right)^{(1/2 + 2\epsilon)n} \cdot \left(\frac{1}{2} - 2\epsilon\right)^{(1/2 - 2\epsilon)n} \geq 2^{-cn\epsilon^2}.$$

Now, using independence between different $w \in W$, we can conclude that, for $n \leq d/(6c\epsilon^2)$, the probability that there exists w for which uniform convergence is violated is at least

$$1 - \left(1 - 2^{-cn\epsilon^2}\right)^{|W|} \geq 1 - e^{-2^{-cn\epsilon^2} \cdot 2^{d/6}} \geq 1 - e^{-1} > \frac{1}{2}.$$

$\square$

A natural question is whether the d/ϵ^2 dependence also holds for ERM. We could not answer it and prove only a weaker $\Omega(d/\epsilon)$ lower bound. For completeness, we also make this statement for general radius R and Lipschitz bound L.

Theorem 3.10. *For $L, R > 0$ and $\epsilon \in (0, LR/4)$, let $\mathcal{K} = \mathcal{B}_2^d(R)$ and we define $\mathcal{H}_2 \doteq \{L \cdot g_V \mid V \subseteq W\} \subseteq \mathcal{F}^0(\mathcal{B}_2^d(R), L)$ for g_V defined in eq. (1). We define the random variable $\mathbf{V}_\alpha$ as a random subset of W obtained by including each element of W with probability $\alpha \doteq 2\epsilon/(LR)$ randomly and independently. Let D_α be the probability distribution of the random variable $g_{\mathbf{V}_\alpha}$. Then for $n \leq d/32 \cdot LR/\epsilon$ and every set of samples S there exists an ERM $\hat{x}(S)$ such that*

$$\Pr_{\mathbf{S} \sim D_\alpha^n} [F_{D_\alpha}(\hat{x}(\mathbf{S})) - F^* \geq \epsilon] > 1/2.$$

Proof. By the same argument as in the proof of Thm. 3.3 we have that: For every $w \in W$, and $V \subseteq W$, $L \cdot g_V(R\bar{w}) = LR$ if $w \in V$ and $L \cdot g_V(R\bar{w}) = LR/2$ otherwise. For $g_{\mathbf{V}}$ chosen randomly with respect to D_α, we have that $w \in \mathbf{V}$ with probability $2\epsilon/(LR)$. This implies that $F_{D_\alpha}(R\bar{w}) = LR/2 + \epsilon$. Similarly, $\min_{\mathcal{K}}(F_{\mathbf{S}}) = LR/2$ and $F^* = \min_{\mathcal{K}}(F_{D_\alpha}) = LR/2$.

Therefore, if $\bigcup_{i \in [n]} \mathbf{V}_i \neq W$ then there exists an ERM $\hat{x}(\mathbf{S})$ such that $F_{\mathbf{S}}(\hat{x}(\mathbf{S})) = \min_{\mathcal{K}}(F_{\mathbf{S}})$ and $F_{D_\alpha}(\hat{x}(\mathbf{S})) - F^* = \epsilon$. For the distribution D_α and every $w \in W$,

$$\Pr_{\mathbf{S} \sim D_\alpha^n} \left[w \in \bigcup_{i \in [n]} \mathbf{V}_i \right] = 1 - (1 - \alpha)^n \leq 1 - e^{-2\alpha n}$$

and this event is independent from the inclusion of other elements in $\bigcup_{i \in [n]} \mathbf{V}_i$ (where we used that $1 - \alpha \geq e^{-2\alpha}$ for $\alpha < 1/2$). Therefore

$$\Pr_{\mathbf{S} \sim D_\alpha^n} \left[\bigcup_{i \in [n]} \mathbf{V}_i = W \right] = \left(1 - e^{-2\alpha n}\right)^{|W|} \leq e^{-e^{-2\alpha n} \cdot e^{d/8}} \leq e^{-1} < \frac{1}{2}.$$

$\square$

4 Lower Bound for Low-Complexity Functions

We will now demonstrate that our lower bounds hold even if one restricts the attention to functions that can be computed efficiently (in time polynomial in d). For this purpose we will rely on known constructions of binary linear error-correcting codes. We describe the construction for non-smooth ℓ_2/ℓ_2 setting but analogous versions of other constructions can be obtained in the same way.

We start by briefly providing the necessary background about binary codes. For two vectors $w^1, w^2 \in \{\pm 1\}^d$ let $\#_{\neq}(w^1, w^2)$ denote the Hamming distance between the two vectors. We say that a mapping $G : \{\pm 1\}^k \to \{\pm 1\}^d$ is a $[d, k, r, T]$ binary error-correcting code if G has distance at least $2r + 1$, G can be computed in time T and there exists an algorithm that for every $w \in \{\pm 1\}^d$ such that for some $z \in \{\pm 1\}^k$, $\#_{\neq}(w, G(z)) \leq r$ finds such z in time T (note that such z is unique).

Given $[d, k, r, T]$ code G, for every $j \in [k]$, we define a function

$$g_j(x) \doteq \max \left\{ 1 - \frac{r}{2d}, \max_{w \in W_j} \langle \bar{w}, x \rangle \right\}, \tag{4}$$

where $W_j \doteq \{G(z) \mid z \in \{\pm 1\}^k, z_j = 1\}$. As before, we note that g_j is convex and 1-Lipschitz (relative to ℓ_2).

Theorem 4.1. *Let G be a $[d, k, r, T]$ code. Let $\mathcal{K} = \mathcal{B}_2^d$ and we define $\mathcal{H}_G \doteq \{g_j \mid j \in [k]\}$ for g_j defined in eq. (4). Let D be the uniform distribution over $\mathcal{H}_G$. Then for every $x \in \mathcal{K}$, $g_j(x)$ can be computed in time $2T + O(d)$. Further, for $n \leq k/2$ and every set of samples $S \in \mathcal{H}_G^n$ there exists an ERM $\hat{x}(S)$ such that*

$$F_D(\hat{x}(S)) - F^* \geq r/(4d).$$

Proof. Let $W \doteq \{G(z) \mid z \in \{\pm 1\}^k\}$. For every distinct $w^1, w^2 \in W$, $\langle \bar{w}^1, \bar{w}^2 \rangle = 1 - 2 \cdot \#_{\neq}(w^1, w^2)/d \leq 1 - (2r + 1)/d < 1 - r/(2d)$. Therefore, by the definition of g_j, for every $w \in W$, $g_j(\bar{w}) = 1$ if $w \in W_j$ and $g_j(\bar{w}) = 1 - r/(2d)$ otherwise. Now for $z \in \{\pm 1\}^k$, let $\#_1(z)$ denote the number of indices $j \in [k]$, such where $z_j = 1$. For $w = G(z)$ there are exactly $\#_1(z)$ indices j, such that $w \in W_j$. This means that for a function g_j chosen randomly with respect to D, we have that $g_j(\bar{w}) = 1$ with probability exactly $\#_1(z)/k$. This implies that $F_D(\bar{w}) = \#_1(z)/k + (1 - \#_1(z)/k)(1 - r/(2d)) = 1 - (1 - \#_1(z)/k) \cdot r/(2d)$.

Let $S = (g_{j_1}, \ldots, g_{j_n})$ be any set of n points from $\mathcal{H}_G$. Observe that $\min_{\mathcal{K}}(F_S) = 1 - r/(2d)$ and $F^* = \min_{\mathcal{K}}(F_D) = 1 - r/(2d)$ (the minimum is achieved at the origin $\bar{0}$). Now, for $I = \{j_1, \ldots, j_n\}$ let z^I denote the vector such that $z_j^I = -1$ if $j \in I$ and $z_j^I = 1$, otherwise. Clearly, $\#_1(z^I) = k - |I| \geq k - n \geq k/2$. Let $w^I \doteq G(z^I)$ and let $\hat{x}(S) \doteq \bar{w}^I$.

Observe that for all $i \in [n]$, $z_{j_i}^I = -1$ and therefore $g_{j_i}(\hat{x}(S)) = 1 - r/(2d)$. This means that $F_S(\hat{x}(S)) = 1 - r/(2d)$ and therefore $\hat{x}(S)$ is a minimizer of F_S. On the other hand, $F_D(\hat{x}(S)) = 1 - (1 - \#_1(z^I)/k) \cdot r/(2d) \geq 1 - r/(4d)$. This implies the claimed generalization error of $r/(4d)$.

Finally we need to show that for any $x \in \mathcal{K}$, $g_j(x)$ can be computed in time $2T + O(d)$. We use the following simple algorithm, let $\text{sign}(x)$ denote the element-wise application of the sign function. We apply the decoding algorithm for G to $\text{sign}(x)$ and z denote the result. If the decoding, succeeds, $z_j = 1$ and $\langle x, \overline{G(z)} \rangle \geq 1 - r/(2d)$, then output $g_j(x) = \langle x, \overline{G(z)} \rangle$. Otherwise, output $g_j(x) = 1 - r/(2d)$. It is easy to see that the running time of this algorithm is $2T + O(d)$.

To prove the correctness, observe that we only need to output a value that is different from $1 - r/(2d)$ when there exists w such that $w \in W_j$ and $\langle x, \bar{w} \rangle > 1 - r/(2d)$. If $\langle x, \bar{w} \rangle > 1 - r/(2d)$ then

$$\|x - \bar{w}\|^2 = \|x\|^2 + \|\bar{w}\|^2 - 2\langle x, \bar{w} \rangle \leq 2 - 2(1 - r/(2d)) = r/d. \tag{5}$$

Now for every $i \in [d]$ such that $\mathsf{sign}(x_i) \neq w_i$, we have that $(x_i - \bar{w}_i)^2 \geq 1/d$. Therefore

$$\|x - \bar{w}\|^2 = \sum_{i \in [d]} (x_i - \bar{w}_i)^2 \geq \frac{\#_{\neq}(\mathsf{sign}(x), w)}{d}.$$

Combining it with eq.(5) we obtain that we only need to output value that is different from $1 - r/(2d)$ only when there exists $w \in W_j$ such that $\#_{\neq}(\mathsf{sign}(x), w) \leq r$. By the properties of G, in this case there is an algorithm that in time T will find the unique z such that $G(z) = w$. Given such z we can compute $G(z)$ in time T and verify that $z_j = 1$ and $\langle x, \overline{G(z)} \rangle \geq 1 - r/(2d)$ in time $O(d)$. $\qquad\square$

We can now use any existing constructions of efficient binary error-correcting codes to obtain a lower bound that uses only a small set of efficiently computable convex functions. Getting a lower bound that has asymptotically optimal dependence on d requires that $k = \Omega(d)$ and $r = \Omega(d)$ (referred to as being *asymptotically good*). The existence of efficiently computable and asymptotically good binary error-correcting codes was first shown by Justesen [Jus72]. More recent work of Spielman [Spi96] shows existence of asymptotically good codes that can be encoded and decoded in $O(d)$ time. In particular, for some constant $\rho > 0$, there exists a $[d, d/2, \rho \cdot d, O(d)]$ binary error-correcting code. As a corollary we obtain the following lower bound.

Corollary 4.2. *Let G be an asymptotically-good $[d, d/2, \rho \cdot d, O(d)]$ error-correcting code for a constant $\rho > 0$. Let $\mathcal{K} = \mathcal{B}_2^d$ and we define $\mathcal{H}_G \doteq \{g_j \mid j \in [d/2]\}$ for g_j defined in eq. (4). Let D be the uniform distribution over $\mathcal{H}_G$. Then for every $x \in \mathcal{K}$, $g_j(x)$ can be computed in time $O(d)$. Further, for $n \leq d/4$ and every set of samples $S \in \mathcal{H}_G^n$ there exists an ERM $\hat{x}(S)$ such that*

$$F_D(\hat{x}(S)) - F^* \geq \rho/4.$$

5   Bounded-Range Convex Optimization

As we have outlined in the introduction, SCO is solvable in the more general setting in which instead of the Lipschitz bound and radius of $\mathcal{K}$ we have a bound on the range of functions in the support of distribution. Recall that for a bound on the absolute value B we denote this class of functions by $\mathcal{F}(\mathcal{K}, B)$. This setting is more challenging algorithmically and has not been studied extensively. For comparison purposes and completeness, we state a recent result for this setting from [RS15] (converted from the online to the stochastic setting in the standard way).

Theorem 5.1 ([RS15]). *Let $\mathcal{K} = \mathcal{B}_2^d(R)$ for some $R > 0$ and $B > 0$. There is an efficient algorithm $\mathcal{A}$ that given $\epsilon, \delta > 0$ and $n = O(\log(B/\epsilon) \log(1/\delta) B^2/\epsilon^2))$ i.i.d. samples from any distribution D supported on $\mathcal{F}(\mathcal{K}, B)$ outputs an ϵ-optimal solution to F_D over $\mathcal{K}$ with probability $\geq 1 - \delta$.*

The case of general $\mathcal{K}$ can be handled by generalizing the approach in [RS15] or using the algorithms in [BLNR15, FGV15]. Note that for those algorithms the sample complexity will have a polynomial dependence on d (which is unavoidable in this general setting).

In contrast to these results, we will now demonstrate that for such problems an ERM algorithm will require an infinite number of samples to succeed already for $d = 2$. As in the proof of Thm. 3.7 we define $f_V(x) = \sum_{w \in V} \phi(\langle w, x \rangle)$. However we can now use the lack of bounds on the Lipschitz constant (or smoothness) to use $\phi(a)$ that is equal to 0 for $a \leq 1 - \alpha$ and $\phi(1) = 1$. For every $m \geq 2$, we can choose a set of m vectors W evenly spaced on the unit circle such that for a sufficiently small $\alpha > 0$, $\phi(\langle w, x \rangle)$ will

not interact with $\phi(\langle w', x \rangle)$, for any two distinct $w, w' \in W$. More formally, let m be any positive integer, let $w^i \doteq (\sin(2\pi \cdot i/m), \cos(2\pi \cdot i/m))$ and let $W_m \doteq \{w^i \mid i \in [m]\}$. Let

$$\phi_\alpha(a) \doteq \begin{cases} 0 & \text{if } a \leq 1 - \alpha \\ (a - 1 + \alpha)/\alpha & \text{otherwise.} \end{cases}$$

For $V \subseteq W_m$ we set $\alpha \doteq 2/m^2$ and define

$$f_V(x) \doteq \sum_{w \in V} \phi_\alpha(\langle w, x \rangle). \tag{6}$$

See Figure 3 for an illustration. It is easy to see that f_V is convex. We now verify that the range of f_V is $[0, 1]$ on $\mathcal{B}_2^2$. Clearly, for any unit vector $w^i \in W_m$, and $x \in \mathcal{B}_2^2$, $\langle w^i, x \rangle \in [-1, 1]$ and therefore $\phi_\alpha(\langle w^i, x \rangle) \in [0, 1]$. Now it suffices to establish that for every $x \in \mathcal{B}_2^2$, there exists at most one vector $w \in W_m$ such that $\phi_\alpha(\langle w, x \rangle) > 0$. To see this, as in Lemma 3.6, we note that if $\phi_\alpha(\langle w, x \rangle) > 0$ then $\langle w, x \rangle > 1 - \alpha$. For $w \in W_m$ and $x \in \mathcal{B}_2^2$, this implies that $\|w - x\| < \sqrt{1 + 1 - 2(1 - \alpha)} = \sqrt{2\alpha}$. For our choice of $\alpha = 2/m^2$, this implies that $\|w - x\| < 2/m$. On the other hand, for $i \neq j \in [m]$, we have

$$\|w^i - w^j\| \geq \|w^1 - w^m\| \geq \sin(2\pi/m) \geq 2\pi/m - (2\pi/m)^3/6 \geq 4/m.$$

Therefore there does not exist x such that $\phi_\alpha(\langle w^i, x \rangle) > 0$ and $\phi_\alpha(\langle w^j, x \rangle) > 0$. Now we can easily establish the lower bound.

Theorem 5.2. *Let $\mathcal{K} = \mathcal{B}_2^2$ and $m \geq 2$ be an integer. We define $\mathcal{H}_m \doteq \{f_V \mid V \subseteq W\}$ for f_V defined in eq. (6). Let D_m be the uniform distribution over $\mathcal{H}_m$. Then for $n \leq \log m$ and every set of samples S there exists an ERM $\hat{x}(S)$ such that*

$$\Pr_{\mathbf{S} \sim D_m^n} \left[F_D(\hat{x}(\mathbf{S})) - F^* \geq 1/2 \right] > 1/2.$$

Proof. Let $\mathbf{S} = (f_{\mathbf{V}_1}, \dots, f_{\mathbf{V}_n})$ be the random samples. Clearly, $F^* = 0$ and $\min_{\mathcal{K}}(F_{\mathbf{S}}) = 0$. Further, the analysis above implies that for every $w \in W_m$ and $V \subseteq W_m$, $f_V(w) = 1$ if $w \in V$ and $f_V(w) = 0$ otherwise. Hence $F_{D_m}(w) = 1/2$. Now, if $\bigcup_{i \in [n]} \mathbf{V}_i \neq W_m$ then let $\hat{x}(\mathbf{S}) \doteq w$ for any $w \in W_m \backslash \bigcup_{i \in [n]} \mathbf{V}_i$. Then for all i, $h_{\mathbf{V}_i}(\hat{x}(\mathbf{S})) = 0$ and hence $F_{\mathbf{S}}(\hat{x}(\mathbf{S})) = 0$. This means that $\hat{x}(\mathbf{S})$ is a minimizer of $F_{\mathbf{S}}$ and $F_{D_m}(\hat{x}(\mathbf{S})) - F^* = 1/2$.

Now, exactly as in Thm. 3.3, we can conclude that $\bigcup_{i \in [n]} \mathbf{V}_i = W_m$ with probability at most

$$\left(1 - 2^{-n}\right)^m \leq e^{-2^{-n} \cdot m} \leq e^{-1} < \frac{1}{2}.$$

$\square$

This lower bound holds for every m. This implies that the sample complexity of $1/2$-optimizing $\mathcal{F}(\mathcal{B}_2^2, 1)$ over $\mathcal{B}_2^2$ is infinite.

6 Discussion

Our work points out to substantial limitations of the classic approach to understanding and analysis of generalization in the context of general SCO. Further, it implies that in order to understand how well solutions

produced by an optimization algorithm generalize, it is necessary to examine the optimization algorithm itself. This is a challenging task that we still have relatively few tools to address. Yet such understanding is also crucial for developing theory to guide the design of optimization algorithms that are used in machine learning applications.

One way to bypass our lower bounds is to use additional structural assumptions. For example, for generalized linear regression problems uniform convergence gives nearly optimal bounds on sample complexity [KST08]. One natural question is whether there exist more general classes of functions that capture most of the practically relevant SCO problems and enjoy dimension-independent (or, scaling as $\log d$) uniform convergence bounds.

An alternative approach is to bypass uniform convergence (and possibly also ERM) altogether. Among a large number of techniques that have been developed for ensuring generalization, the most general ones are based on notions of stability [BE02, SSSSS10]. However, known analyses based on stability often do not provide the strongest known generalization guarantees (*e.g.* high probability bounds require very strong assumptions). Another issue is that we lack general algorithmic tools for ensuring stability of the output. Therefore many open problems remain and significant progress is required to obtain a more comprehensive understanding of this approach. Some encouraging new developments in this area are the use of notions of stability derived from differential privacy [DFH$^+$14, DFH$^+$15, BNS$^+$16] and the use of techniques for analysis of convergence of convex optimization algorithms for proving stability [HRS16].

Acknowledgements

References

[BE02]   Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

[BLNR15]   Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *COLT*, pages 240–265, 2015.

[BM13]   Francis R. Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *NIPS*, pages 773–781, 2013.

[BNS$^+$16]   Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059, 2016.

[Bub15]   Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[CCG04]   N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[DFH$^+$14]   Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.

[DFH+15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *CoRR*, abs/1506, 2015. Extended abstract in NIPS 2015.

[FGV15] Vitaly Feldman, Cristobal Guzman, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. *CoRR*, abs/1512.09170, 2015. Extended abstract in SODA 2017.

[HRS16] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, pages 1225–1234, 2016.

[Jus72] J. Justesen. Class of constructive asymptotically good algebraic codes. *IEEE Trans. Inf. Theor.*, 18(5):652 – 656, 1972.

[KST08] S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, pages 793–800, 2008.

[NJLS09] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

[RS15] Alexander Rakhlin and Karthik Sridharan. Sequential probability assignment with binary alphabets and large classes of experts. *CoRR*, abs/1501.07340, 2015.

[RSS12] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, 2012.

[SN05] A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In V. Jeyakumar and A. M. Rubinov, editors, *Continuous Optimization: Current Trends and Applications 144*. Springer, 2005.

[Spi96] D. Spielman. Linear-time encodable and decodable error-correcting codes. *IEEE Transactions on Information Theory*, 42(6):1723–1731, 1996.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[SSSS09] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[SSSSS10] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[SZ13] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *ICML*, pages 71–79, 2013.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.