Stochastic Convex Optimization: Learnability, Stability and Gradient Descent

Genghis Luo KL4747@NYU.EDU
NYU Shanghai
Nikolaos Tsilivis NT2231@NYU.EDU
NYU

Editor: Mehryar Mohri

Abstract

We consider learning in the general framework of stochastic convex optimization (SCO). First, we study fundamental questions in this area: When is a SCO problem learnable? Is empirical risk minimization enough to guarantee learnability in SCO, as in binary classification? What is the role of stability in learning algorithms for SCO? We cover the perhaps surprising answers to these questions as provided by Shalev-Shwartz et al. (2010).

We then turn our attention to gradient descent (GD) and study its power as a learning algorithm within SCO. A recent result by Schliserman et al. (2024) shows that there are instances where GD requires $\Omega(\sqrt{d})$ samples to arrive at a solution of non-trivial test error (d is the input dimension). A further improvement provided by Livni (2024) establishes that this lower bound is in fact $\Omega(d)$. Taken together, the results imply that there is no statistical benefit of GD in worst-case SCO.

Keywords: Stochastic Convex Optimization, Empirical Risk Minimization, Gradient Descent

1. Introduction

Recent advances in machine learning have equated, in the minds of most, learning with classification (in the case of pattern recognition applications, e.g., medical prognoses) or density estimation (in the case of so-called generative AI, e.g., large language/diffusion models). However, learning can be defined in its most abstract form with respect to any loss function. Formally, Vapnik (1995) defined the *General Setting of Learning*. In the General Setting of Learning, a learning problem $\mathcal{P}$ is a tuple $(\mathcal{Z}, \mathcal{D}, \mathcal{H}, f)$ where:

- $\mathcal{Z}$ is a measurable instance set.
- $\mathcal{D}$ is a distribution over the instance set $\mathcal{Z}$.
- $\mathcal{H}$ is a hypothesis set.
- $f : \mathcal{H} \times \mathcal{Z} \mapsto \mathbb{R}$ is a measurable loss function.

The goal of learning is to minimize the *population loss/risk*:

$$\min_{h \in \mathcal{H}} F(h) := \mathbb{E}_{z \sim \mathcal{D}}[f(h, z)] \tag{1}$$

That is, we would like to select an h belonging in $\mathcal{H}$ whose risk is as close as possible to the minimum value $\inf_{h \in H} F(h)$. But since the distribution $\mathcal{D}$ of $\mathcal{Z}$ is unknown, we cannot do this directly, but instead need to rely on finite samples which are often assumed to be i.i.d, i.e.,

$S = \{z_1, \dots, z_m\} \sim \mathcal{D}^m$. For these samples, we denote the *empirical loss/risk* by $\hat{F}(h) = \frac{1}{m}\sum_{i=1}^m f(h, z_i)$. On this sample, we apply a learning rule $\mathcal{A} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ to pick a hypothesis.

This report focuses on one particular instantiation of the above framework: *Stochastic Convex Optimization (SCO)*. In SCO, we impose the following further limitations to the learning problems:

- The hypothesis set $\mathcal{H}$ is a closed, convex and bounded subset of a Hilbert space.

- The loss function $f : \mathcal{H} \times \mathcal{Z} \mapsto \mathbb{R}$ is Lipschitz-continuous and convex w.r.t. its first argument.

That is, learning is restricted to convex spaces. Note that this setup differs from familiar binary classification, for example, as there success of learning is measured with respect to the zero-one loss $f(h, z = (x, y)) = \mathbb{1}\{h(x) \neq y\}$, and the hypothesis set $\mathcal{H}$ may or may not be convex. However, we can have supervised learning problems framed as SCO instances.

Example 1 *Let instance space $\mathcal{Z} = \mathcal{X} \times \{\pm 1\}$ with $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq B\}$, $B > 0$, hypothesis space $\mathcal{H} = \{h \in \mathbb{R}^d : \|h\|_2 \leq W\}$, $W > 0$ and let the loss function to be defined as $f(h, (x, y)) = l(\langle x, h\rangle, y)$ for some convex and Lipschitz loss function l.*

In this example, we know that uniform convergence holds: for any distribution $\mathcal{D}$, with high probability over $z_1, \dots, z_m \sim \mathcal{D}$:

$$\sup_{h \in \mathcal{H}} \left| F(h) - \hat{F}(h) \right| \xrightarrow{m \to \infty} 0. \tag{2}$$

This justifies selecting an empirical risk minimizer (ERM) as a predictor:

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{F}(h), \tag{3}$$

as its population loss $F(\hat{h})$ is guaranteed to converge to the optimal population loss $F(h^\star) = \inf_{h \in \mathcal{H}} F(h)$ as $m \to \infty$.

What about general SCO problems? Can we always select an ERM, with the confidence that its population loss will converge to the optimality? Let us formally define this desideratum to make the discussion smoother going forward.

Definition 1 (Learnability): A learning problem $\mathcal{P} = (\mathcal{Z}, \mathcal{D}, \mathcal{H}, f)$ is *learnable* if there exists a learning rule $\mathcal{A}$ and a monotonically decreasing sequence $\epsilon_{\text{cons}}(m)$, such that $\epsilon_{\text{cons}}(m) \xrightarrow{m \to \infty} 0$, and, for all distributions $\mathcal{D}$, it holds:

$$\mathbb{E}_{S \sim \mathcal{D}^m} \left[F(\mathcal{A}(S)) - \inf_{h \in \mathcal{H}} F(h) \right] \leq \epsilon_{\text{cons}}(m). \tag{4}$$

A learning rule that satisfies the above is called *universally* consistent.

Remark 1 *The term consistent refers to the requirement that eq. (4) holds for all distributions $\mathcal{D}$. This is perhaps one place where the definition of learnability becomes too strong to be useful in applications. However, this is consistent with the usual definition of PAC-learnability (which additionally brings up computational considerations).*

Let us restate the main questions we are interested in understanding SCO problems: Under what conditions is an SCO problem $\mathcal{P}$ learnable? Is ERM the only rule we need to consider to determine learnability?

To kick things off, let us provide the answer right away for **finite-dimensional** learning problems.

Theorem 2 (Shalev-Shwartz et al. (2009)) *Let $\mathcal{P} = (\mathcal{Z}, \mathcal{D}, \mathcal{H}, f)$ be a stochastic convex optimization problem where $\mathcal{H} \subset \mathbb{R}^d$ is bounded by B and let $f(h, z)$ be L-Lipschitz w.r.t. h. Then with probability of at least $1 - \delta$ over a sample of size m, for all $h \in \mathcal{H}$, we have:*

$$|F(h) - \hat{F}(h)| \leq \mathcal{O}\left(\sqrt{\frac{L^2 B^2 d \log(m) \log\left(\frac{d}{\delta}\right)}{m}}\right). \tag{5}$$

This result is a messenger of both good and bad news. On one hand, the above theorem shows that the population loss converges uniformly to the empirical loss as m goes to ∞, which means that any finite-dimensional problem $\mathcal{P}$ with a Lipschitz loss f is learnable via ERM. On the other hand, the dependence on d is alarming: it indicates that in high dimensions, there might be a high price to pay for generalization. It also permits the possibility of failure of uniform convergence in infinite dimensions.

This is, in fact, the case (Shalev-Shwartz et al., 2010). As we shall see in Section 2, there are SCO learning problems in infinite dimensions where uniform convergence does not hold, ERM fails to output a generalizable solution, yet the problem is still learnable via alternative learning rules. The rest of that section is devoted to characterizing what alternative rules are guaranteed to work for learnable problems. The key notion there is that of the *stability* of a learning algorithm.

In Section 3, we return to finite-dimensional learning problems and present a dimension-dependent lower bound on the number of samples required for an ERM to generalize in SCO, which is due to Feldman (2016).

Despite this negative result, there is hope that common algorithms (specific learning rules, which may either be special ERMs or may not minimize the empirical risk at all (Zinkevich, 2003)) may not require such pessimistic dimension-dependent sample complexities. In two recent papers, Schliserman et al. (2024) and Livni (2024) settle this question: they show that gradient descent (GD) and (unstable) stochastic gradient descent (SGD), the prototypical first-order methods used in SCO and beyond, also suffer from the same dimension-dependent lower bounds. This is bad news because their results imply that the theoretical tools we possess are unable to differentiate between ERMs and solutions returned by algorithms people use in practice (where we know that a gap between the two exists (Zhang et al., 2017)). These results are presented in Section 4.

Finally, in Section 5, we outline some directions worth studying in the future.

2. Learnability in Stochastic Convex Optimization

As we mentioned in the introduction, there exists a simple stochastic convex optimization problem that is not learnable with ERM, despite being learnable otherwise. We proceed with the construction which appears in Shalev-Shwartz et al. (2010):

Example 2 *Let $\mathcal{H}$ be the unit sphere of an infinite-dimensional Hilbert space with orthonormal basis $e_1, e_2, \ldots$ and let $\mathcal{Z} = \mathcal{H} \times V$, where V is an infinite-dimensional Hilbert space with each*

coordinate belonging to $[0,1]$. *Let the loss function be:*

$$f(h,(x,\alpha)) = \|\alpha \cdot (h-x)\|. \tag{6}$$

The distribution $\mathcal{D}$ *is specified as follows:* $x = 0$ *with probability 1 and each* α_i *is an independent Bernoulli random variable. Suppose we have* m *samples from this distribution:* $S = \{(x^{(1)},\alpha^{(1)}),\dots,(x^{(m)},\alpha^{(m)})\}$. *Then, there exists almost surely an index* $j^\star$ *such that* $\alpha_j^{(i)} = 0$ *for all* $i \in [m]$. *As a result, the basis vector* $e_{j^\star}$ *is an ERM:*

$$\hat{F}(e_j) = \frac{1}{m}\sum_{i=1}^{m} \|\alpha \cdot (e_j - 0)\| = 0. \tag{7}$$

However, the population loss of this ERM is poor:

$$F(e_j) = \mathbb{E}_{(x,\alpha)\sim\mathcal{D}}\left[\alpha \cdot (e_j - 0)\right] = \frac{1}{2}. \tag{8}$$

Thus, the deviation of the population from the empirical loss can be arbitrarily bad, even for ERMs:

$$\sup_{h\in\mathcal{H}} \left|F(h) - \hat{F}(h)\right| \geq \frac{1}{2}. \tag{9}$$

This demonstrates that uniform convergence does not hold in this case and, furthermore, ERM is not guaranteed to generalize well. Note that the zero predictor[1] $h = 0$ *has the optimal population loss* $F(0) = 0$.

The conclusion from the above elementary construction is remarkably surprising: uniform convergence may not hold in SCO and ERM is not guaranteed to generalize well! On the contrary, problems such as the above are learnable via alternative learning rules that have nothing to do with uniform convergence. The key element is to introduce *stability* or, equivalently, *regularization*.

Theorem 3 (Shalev-Shwartz et al. (2010)) *Let* $\mathcal{P} = (\mathcal{Z},\mathcal{D},\mathcal{H},f)$ *be a stochastic convex optimization problem where* $f : \mathcal{H}\times\mathcal{Z} \to \mathbb{R}$ *is L-Lipschitz with respect to* h *and* $\mathcal{H}$ *is bounded by* B. *Let* $z_1,\dots,z_m$ *be an i.i.d. sample and let* $\hat{h}_\lambda$ *be defined as:*

$$\hat{h}_\lambda = \arg\min_{h\in\mathcal{H}} \left(\frac{1}{m}\sum_{i=1}^{m} f\left(h,z_i\right) + \frac{\lambda}{2}\|h\|^2 \right), \tag{10}$$

for $\lambda = \sqrt{\frac{16L^2}{\delta B^2 m}}$. *Then, with probability at least* $1-\delta$ *we have:*

$$F\left(\hat{h}_\lambda\right) - \inf_{h\in\mathcal{H}} F\left(h\right) \leq \sqrt{\frac{8L^2B^2}{\delta m}}. \tag{11}$$

That is, regularized empirical risk minimization (RERM) is guaranteed to generalize well, even though the solution returned by vanilla ERM might be arbitrarily poor. The authors of Shalev-Shwartz et al. (2010) discuss how regularization here plays a different role than in traditional learning theory: the role of regularization is to make the learning algorithm stable and **NOT** to control the capacity of the hypothesis class.

1. Note that $h = 0$ is also an ERM. The above example can be slightly modified to yield an example where the unique ERM generalizes poorly.

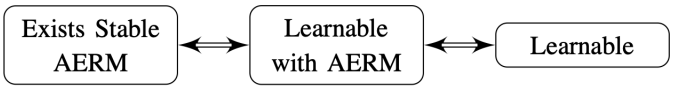

Figure 1: Learnability in the General Setting of Vapnik – credit to Shalev-Shwartz et al. (2010).

Remark 4 *In addition to regularized ERM, we also know an alternative learning rule which is guaranteed to perform well in SCO problems: Zinkevich's (Zinkevich, 2003) online learning algorithm together with an online-to-batch conversion. The regularization there is implicit: Zinkevich's algorithm can be viewed as approximate coordinate ascent optimization of the dual of the regularized problem (10).*

The proof of Theorem 3 crucially depends on the notion of stability (regularized ERM is shown to be stable). We review the standard definition of (uniform) stability of a learning algorithm.

Definition 2 (Uniform Stability): Let S and S' be any two training samples that differ by a single point. A learning algorithm $\mathcal{A} : \cup_{m=1}^{\infty} \mathcal{Z}^m \mapsto \mathcal{H}$ is said to be *uniformly β-stable* if the hypotheses it returns when trained on S and S' satisfy

$$\forall z \in \mathcal{Z}, \quad |f(h_S, z) - f(h_{S'}, z)| \leq \beta. \tag{12}$$

The smallest such β satisfying this inequality is called the *stability coefficient* of $\mathcal{A}$.

Remarkably, the above discussion extends to any problem in the General Setting of Learning. To state (in passing) the result, we need the following definition of *asymptotic ERM* (AERM).

Definition 3 (AERM): We say that a learning rule $\mathcal{A} : \cup_{m=1}^{\infty} \mathcal{Z}^m \mapsto \mathcal{W}$ is an *asymptotic ERM* with rate $\epsilon_{\mathrm{ERM}}(m)$ under $\mathcal{D}$ if:

$$\mathbb{E}_{S \sim \mathcal{D}^m} \left[\hat{F}(\mathcal{A}(S)) - \hat{F}(\hat{w}) \right] \leq \epsilon_{\mathrm{ERM}}(m). \tag{13}$$

A learning rule is *universally* an AERM with rate $\epsilon_{\mathrm{ERM}}(m)$, if it is an AERM with rate $\epsilon_{\mathrm{ERM}}(m)$ for each distribution $\mathcal{D}$.

Remarkably, Shalev-Shwartz et al. (2010) are able to provide an exact characterization of learnability.

Theorem 5 (Informal- Shalev-Shwartz et al. (2010)) *A learning problem is learnable if and only if there exists a stable, universally AERM learning rule.*

Thus, for general learning problems, it suffices to search over asymptotic ERM learning rules – see Figure 1.

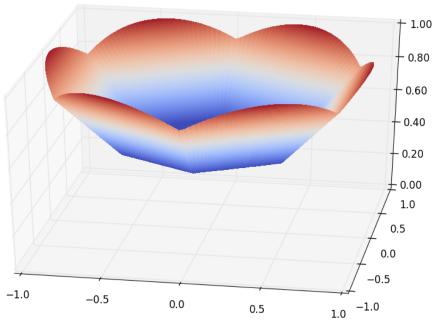

Figure 2: The Feldman (2016) construction in 2 dimensions.

3. Sample Complexity of ERM

We, now, return to finite-dimensional stochastic convex optimization problems which abound in applications. In this section, we construct a learning problem in high-dimensions where an ERM generalizes poorly. We then discuss how this translates to an $\Omega(d)$ lower bound on the sample complexity of ERM in SCO. These contributions appear in Feldman (2016).

The construction depends on the probabilistic method, first pioneered by Paul Erdos.

Example 3 *Let U be a set of unit, nearly-orthogonal vectors in $\mathbb{R}^d$ such that $|\langle u, v\rangle| \leq 1/8$ for all $u \neq v$ in U. Let $P(U)$ be the powerset of U and let $\mathcal{D}$ be the uniform distribution over $P(U)$. That is, $\mathcal{Z} \subseteq U$. Let $V_1, \dots, V_m \sim \mathcal{D}$ and take $m = d$. The loss function[2] is defined as follows:*

$$f_{F16}(h, V) = \max\left\{\frac{1}{2}, \max_{u \in V}\langle u, w\rangle\right\}. \tag{14}$$

Note that is convex and Lipschitz in its first argument. We compute the probability of not sampling a $u \in U$ in the m V_i's:

$$\begin{aligned}
\mathbb{P}\left[\exists u_0 \in U : \ u_0 \notin V_i, \, i \in [m]\right] &= 1 - \mathbb{P}\left[\forall u_0 \in U : \ \exists i \in [m] \ u_0 \in V_i\right] \\
&= 1 - \prod_{j=1}^{d} \mathbb{P}\left[\exists i : u_j \in V_i\right] \\
&= 1 - \left(1 - \mathbb{P}\left[u_j \notin V_i \, \forall i \in [m]\right]\right)^d \\
&= 1 - \left(1 - \frac{1}{2^d}\right)^{2^d} \\
&\geq 1 - e.
\end{aligned} \tag{15}$$

For the "unsampled" vector u_0, it holds for all $i \in [m]$:

$$f_{F16}(u_0, V_i) = \frac{1}{2}, \tag{16}$$

2. See Figure 2

hence u_0 is an ERM. However:

$$F_{F16}(u_0) = \mathbb{E}_{V\sim\mathcal{D}}\left[f_{F16}(u_0, V)\right] = \frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} = \frac{3}{4}. \tag{17}$$

Thus, with constant probability:

$$\left|F_{F16}(u_0) - \hat{F}_{F16}(u_0)\right| \geq 1/4 > 0. \tag{18}$$

As a result, ERM generalizes poorly and uniform convergence fails.

Finally, note that our construction assumed the existence of a set U of nearly orthogonal unit vectors. Its existence can be shown via standard probabilistic arguments in dimensions high enough.

We saw an example where the number of dimensions is equal to the number of samples $m = d$ where ERM fails to generalize well. As a corollary, we obtain that, within SCO, ERM requires at least $\Omega(d)$ in order to generalize.

4. Sample Complexity of GD

The previous results permit some optimism: while an arbitrary ERM might not be able to generalize well in high-dimensional SCO problems, the specific properties of learning rules used in practice, such as first-order methods, could perhaps lead them in less pathological solutions. For this reason, we shift our attention to the sample complexity of gradient descent (GD) and stochastic gradient descent (SGD). We present the results of Schliserman et al. (2024).

4.1. Gradient Descent & Stochastic Gradient Descent

First, we introduce the algorithms we consider. In this section, we denote the hypothesis space by $\mathcal{W} \subset \mathbb{R}^d$. We assume access to m samples $z_1, \ldots, z_m$.

Gradient Descent (full-batch)

initialize at $w_1 \in \mathcal{W}$;

update $w_{t+1} = \Pi_{\mathcal{W}}\left(w_t - \frac{\eta}{m}\sum_{i=1}^{m}\nabla_{w_t} f(w_t, z_i)\right),\ 1 \leq t < T$;

return $\bar{w}_S := \frac{1}{S}\sum_{i=1}^{S} w_{T-i+1}$.

where $\Pi_{\mathcal{W}}$ denotes the projection onto the convex set $\mathcal{W} \subset \mathbb{R}^d$, and $\eta > 0$ is the learning rate. The output is the average over the last S iterates.

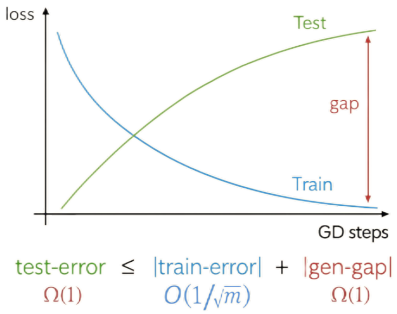

Figure 3: Overfit of GD

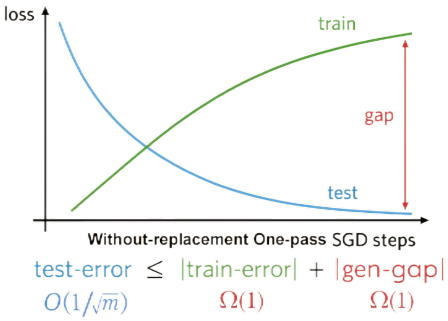

Figure 4: Underfit of SGD

Stochastic Gradient Descent

initialize at $w_1 \in \mathcal{W}$;

update $w_{t+1} = \Pi_{\mathcal{W}}\Big(w_t - \eta \nabla_{w_t} f(w_t, z_i)\Big),\ 1 \le t < T$;

return $\bar{w}_S := \dfrac{1}{S}\sum_{i=1}^{S} w_{T-i+1}$.

4.2. Lower bounds

Recently, Schliserman et al. (2024) showed two important results for these two algorithms – see Figures 3, 4. They provided:

- A construction of a learning problem, where GD (starting from a data-independent initialization) outputs a bad ERM, unless trained with $m = \Omega(\sqrt{d})$ samples.
- A construction of a learning problem, where SGD outputs an *underfit* model, unless trained with $m = \tilde{\Omega}(\sqrt{d})$ samples.

For reference, the previous such known result was of the form $m = \Omega(\log d)$ due to Amir et al. (2021). Thus, the results of Schliserman et al. (2024) provide an exponential improvement of the lower bounds.

We first present the GD lower bound.

Theorem 6 (Schliserman et al. (2024)) *Fix $m > 0$, $T > 3200^2$ and $0 \le \eta \le \frac{1}{5\sqrt{T}}$ and let $d = 178mT + 2m^2 + \max\{1, 25\eta^2T^2\}$. There exists a distribution $\mathcal{D}$ over instance set $\mathcal{Z}$ and a convex, differentiable and 1-Lipschitz loss function $f : \mathbb{R}^d \times Z \to \mathbb{R}$ such that for GD (either projected or unprojected; with $W = \mathbb{B}^d$ or $W = \mathbb{R}^d$ respectively) initialized at $w_1 = 0$ with step size η, for all $t = 1, \dots, T$, the t-suffix averaged iterate has, with probability at least $\frac{1}{6}$ over the choice of the training sample,*

$$F(\bar{w}_{T,t}) - F(w^\star) = \Omega\left(\min\left\{\eta\sqrt{T} + \frac{1}{\eta T}, 1\right\}\right).$$

Proof *(Proof sketch)* The construction is similar to that of Feldman (2016), adjusted in order to make GD progress towards a bad ERM. Let U defined as in the Feldman construction as the set of nearly orthogonal unit vectors, $\mathcal{W} = \mathbb{R}^d, \mathcal{Z} = P(U) \times [m^2]$. The distribution $\mathcal{D}$ is uniform over $P(U) \times [m^2]$. The loss function $f : \mathcal{W} \times \left(P(U) \times [m^2]\right) \mapsto \mathbb{R}$ consists of four different terms:

$$f(w, (V, j)) = l_1(w, V) + l_2(w, (V, j)) + l_3(w) + l_4(w), \tag{19}$$

where:

-
$$l_1(w, V) = \sqrt{\sum_{k=2}^{T} f_{F16}^2(w^{(k)}, V)}, \tag{20}$$

  where $w^{(k)} \in \mathbb{R}^{d'}$ and $f_{F16}(w, V) = \max\left\{\frac{3}{32}\eta, \max_{u \in V}\langle u, w\rangle\right\}$. This term is responsible for the existence of the bad ERM.

-
$$l_2(w, (V, j)) = \langle -\phi(V, j), w^{(0)}\rangle, \tag{21}$$

  where $w^{(0)} \in \mathbb{R}^{2m^2}$ and ϕ "encodes" each V. This term is responsible for encoding the dataset into the weights.

-
$$l_3(w) = \max\left\{\delta_1, \max_{\psi \in \Psi}\left\{\langle \psi, w^{(0)}\rangle - \beta\langle \alpha(\psi), w^{(1)}\rangle\right\}\right\}, \tag{22}$$

  where Ψ encode all the possible datasets and $\alpha(\psi)$ returns a vector $u \in U$ that does not belong to the dataset associated with the encoding ψ. This term is responsible for steering the first component towards the bad ERM.

-
$$l_4(w) = \max\left\{\delta_1, \max_{u \in U, k<T}\left\{\frac{3}{8}\langle u, w^{(k)}\rangle - \frac{1}{2}\langle u, w^{(k+1)}\rangle\right\}\right\} \tag{23}$$

  This final term is responsible for making all the components of the parameter vector $w^{(1)}, \ldots, w^{(k)}$ steer towards the bad ERM.

The proof proceeds to show that GD initially makes one bad step towards the bad ERM in one of the orthogonal subspaces, and then it repeats this bad step in the rest of the subspaces. ■

The implication of the previous result becomes apparent if we plug in the following values as referred to as the *canonical setting*. Let $T = m$ and $\eta = \Theta(1/\sqrt{m})$, then $d = 178mT + 2m^2 + \max\left\{1, 25\eta^2T^2\right\} = \Theta(m^2)$ and it also holds for the generalization gap:

$$F(w_{T,t}) - F(w^\star) = \Omega(1). \tag{24}$$

The implication is that at least $m \geq \Omega(\sqrt{d})$ samples are required for GD to reach nontrivial population loss.

A similar construction establishes a similar lower bound for (non-stable) SGD.

Theorem 7 (Schliserman et al. (2024)) *Fix $m > 2048$ and $0 \leq \eta \leq \frac{1}{5\sqrt{m}}$ and let $d = 712m \log m + 2m^2 + \max\left\{1, 25\eta^2 m^2\right\}$. There exists a distribution $\mathcal{D}$ over instance set $\mathcal{Z}$ and a convex, 1-Lipschitz and differentiable loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ such that for one-pass SGD (either projected or unprojected; with $W = \mathbb{B}^d$ or $W = \mathbb{R}^d$ respectively) over $T = m$ steps initialized at $w_1 = 0$ with step size η, for all $t = 1, \ldots, T$, the t-suffix averaged iterate has, with probability at least $\frac{1}{2}$ over the choice of the training sample,*

$$\hat{F}(w_{T,t}) - \hat{F}(\hat{w}_\star) = \Omega\left(\min\left\{\eta\sqrt{T} + \frac{1}{\eta T}, 1\right\}\right).$$

Remark 8 *It is well known (see, e.g., Hazan (2016), Shalev-Shwartz and Ben-David (2014)) that if one runs SGD with a learning rate $\eta = \Theta(1/\sqrt{m})$ for $T = m$ iterations, then the output w_S^{SGD} holds:*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[F\left(w_S^{\mathrm{SGD}}\right)\right] \leq F(w^*) + \mathcal{O}(1/\sqrt{m}). \tag{25}$$

The above result does not contradict the behavior of SGD in this canonical setting, where it is known that SGD generalizes.

5. Conclusion

In this report, we considered learning in the fundamental framework of Stochastic Convex Optimization and we saw how the learning landscape is surprisingly more complex than in standard supervised learning (Shalev-Shwartz et al., 2010). An ERM might fail arbitrarily, yet the learning problem might still be learnable with alternative, stable, learning rules which nevertheless are approximate ERMs.

Initially, these challenges motivated the study of learning algorithms and first-order methods within SCO to understand when learning problems are easy. However, in a series of recent papers, it has been established that the sample complexity of GD (an archetypic algorithm used in practice) is in worst-case as bad as an arbitrary ERM. A question then remains: **Can we identify a theoretical framework where GD is provably better than an arbitrary ERM?** This is no doubt the most interesting open question we have agreed upon, yet there are more technical questions remain as well. To list a few, is dimensional dependency indeed linear for general full-batch first-order algorithms other than GD and SGD? Do we have an dimensional-dependent upper bound on the optimization/training error of one-pass without-replacement SGD (as it is not an ERM)?

References

I Zaghloul Amir, Tomer Koren, and Roi Livni. Sgd generalizes better than gd (and regularization doesn't help). *ArXiv*, abs/2102.01117, 2021. URL https://api.semanticscholar.org/CorpusID:231749556.

Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/8c01a75941549a705cf7275e41b21f0d-Paper.pdf.

Elad Hazan. Introduction to online convex optimization. *Found. Trends Optim.*, 2(3–4):157–325, August 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL https://doi.org/10.1561/2400000013.

Roi Livni. The sample complexity of gradient descent in stochastic convex optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=2INcTKPBy4.

Matan Schliserman, Uri Sherman, and Tomer Koren. The dimension strikes back with gradients: Generalization of gradient methods in stochastic convex optimization. In *36th International Conference on Algorithmic Learning Theory*, 2024. URL https://openreview.net/forum?id=Y08eqbmCDD.

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

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Annual Conference Computational Learning Theory*, 2009. URL https://api.semanticscholar.org/CorpusID:1016397.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010. URL http://jmlr.org/papers/v11/shalev-shwartz10a.html.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 928–935. AAAI Press, 2003. ISBN 1577351894.