Benign Underfitting of Stochastic Gradient Descent

Tomer Koren
Blavatnik School of Computer Science
Tel Aviv University and Google Research
tkoren@tauex.tau.ac.il

Roi Livni
Department of Electrical Engineering
Tel Aviv University
rlivni@tauex.tau.ac.il

Yishay Mansour
Blavatnik School of Computer Science
Tel Aviv University and Google Research
mansour.yishay@gmail.com

Uri Sherman
Blavatnik School of Computer Science
Tel Aviv University
urisherman@mail.tau.ac.il

Abstract

We study to what extent may stochastic gradient descent (SGD) be understood as a "conventional" learning rule that achieves generalization performance by obtaining a good fit to training data. We consider the fundamental stochastic convex optimization framework, where (one pass, *without*-replacement) SGD is classically known to minimize the population risk at rate $O(1/\sqrt{n})$, and prove that, surprisingly, there exist problem instances where the SGD solution exhibits both empirical risk and generalization gap of $\Omega(1)$. Consequently, it turns out that SGD is not algorithmically stable in *any* sense, and its generalization ability cannot be explained by uniform convergence or any other currently known generalization bound technique for that matter (other than that of its classical analysis). We then continue to analyze the closely related *with*-replacement SGD, for which we show that an analogous phenomenon does not occur and prove that its population risk does in fact converge at the optimal rate. Finally, we interpret our main results in the context of without-replacement SGD for finite-sum convex optimization problems, and derive upper and lower bounds for the multi-epoch regime that significantly improve upon previously known results.

1 Introduction

Conventional wisdom in statistical learning revolves around what is traditionally known as the bias-variance dilemma; the classical theory stipulates the quality of fit to the training data be in a trade-off with model complexity, aiming for a sweet spot where training error is small but yet representative of performance on independent test data.

This perspective is reflected in the vast majority of generalization bound techniques offered by contemporary learning theory. Uniform convergence approaches [35, 4] seek capacity control over the model function class, and employ uniform laws of large numbers to argue convergence of sample averages to their respective expectations. Algorithmic stability [9, 31] on the other hand, builds on controlling sensitivity of the learning algorithm to small changes in its input, and provides algorithm dependent bounds. Nevertheless, despite the conceptual and technical differences between these two methods, both ultimately produce risk bounds by controlling the training error, and the *generalization gap*. The same is true for many other techniques, including sample compression [17, 2], PAC-Bayes [18, 12], and information theoretic generalization bounds [28, 36, 24], to name a few.

In recent years it has become clear there are other, substantially different, ways to manage the fit vs. complexity trade-off, that are in a sense incompatible with traditional generalization bound techniques. Evidently, heavily over-parameterized deep neural networks may be trained to perfectly

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

fit training data and generalize well nonetheless [37, 25], thus seemingly disobeying conventional statistical wisdom. This phenomenon has garnered significant attention, with a flurry of research works dedicated to developing new techniques that would be able to explain strong generalization performance of algorithms in this so called interpolation regime (see [6, 8] and references therein). Notably, while these algorithms do not strike a balance between model complexity and fit to the data in the traditional sense, fundamentally, they still minimize the empirical risk as a proxy to test performance.

To summarize, in the classical and modern regimes alike, learning methods are thought of as minimizing some combination of the training error and generalization gap, with reasoning that relies in one way or another on the following trivial, yet arguably most profound, bound:

$$\text{test-error} \;\leq\; |\text{train-error}| \;+\; |\text{generalization gap}| \,. \tag{1}$$

In this work, we focus on stochastic gradient descent (SGD)—the canonical algorithm for training machine learning models nowadays—and ask whether its generalization performance can be understood through a similar lens. We consider the fundamental stochastic convex optimization (SCO) framework, in which it is well known that SGD minimizes the population risk at a rate of $O(1/\sqrt{n})$ [23]. Remarkably, the classical analysis targets the population risk directly, and in contrast with other generalization arguments, at least seemingly *does not* rely on the above bound. This highlights an intriguing question: Are these quantities, so fundamental to learning theory, relevant to the way that SGD "works"? Put differently, is it possible to provide a more "conventional" analysis of SGD that conforms with (1)?

Our main result shows that, perhaps surprisingly, there exist convex learning problems where the above bound becomes vacuous for SGD: namely, SGD minimizes the population risk, but at the same time, it *does not* minimize the empirical risk and thus exhibits constant generalization gap. This accords neither with the traditional viewpoint nor with that of interpolation, as both recognize the empirical risk as the principal minimization objective. We refer to this phenomenon as *benign underfitting*: evidently, SGD underfits the training data, but its classical analysis affirms this underfitting to be *benign*, in the sense that test performance is never compromised as a result. Our construction presents a learning problem where the output of SGD with step size η over n i.i.d. training examples is $\Omega(\eta\sqrt{n})$ sub-optimal w.r.t. the best fit possible, and consequently has a generalization gap of the same order. Notably, with the standard step size choice of $1/\sqrt{n}$ necessary to ensure the population risk converges at the optimal rate this lower bound amounts to a constant.

Many previously plausible explanations for generalization properties of this algorithm are thereby rendered inadequate, at least in the elementary convex setup we consider here. First, it is clear that SGD cannot be framed as any reasonable regularized empirical risk minimization procedure for the simple reason that it does not minimize the empirical risk, which challenges the implicit regularization viewpoint to the generalization of SGD. Second, any attempt to explain generalization of SGD by uniform convergence over any (possibly data-dependent) hypotheses set cannot hold, simply because the sample average associated with the very same training set SGD was trained on is not necessarily close to its respective expectation. Finally, as it turns out, SGD provides for a strikingly natural example of an algorithm that generalizes well but is not stable in *any* sense, as the most general notion of algorithmic stability is entirely equivalent to the generalization gap [31].

We then move on to study the generalization gap and empirical risk guarantees of SGD in a broader context. We study the case of non-convex and strongly convex component functions, and present natural extensions of our basic result. In addition, we analyse the variant of SGD where datapoints are sampled with-replacement from the training set, in which case the train error is of course low but perhaps surprisingly the population risk is well behaved. Finally, we make the natural connection to the study of without-replacement SGD for empirical risk minimization, and derive upper and lower bounds for the multi-epoch regime. These last two points are discussed in further detail in the following.

With vs without-replacement SGD. We may view one-pass SGD as processing the data via *without*-replacement sampling from the training set, as randomly reshuffling the examples does not change their unconditional distribution. Thus, it is interesting to consider the generalization gap of the closely related algorithm given by running SGD over examples sampled *with*-replacement from the training set. Considering instability (see Appendix A for a detailed discussion) of SGD for non-smooth losses and the fact that this variant targets the empirical objective, a priori it would

seem this algorithm would overfit the training set and not provide strong population risk guarantees. Surprisingly, our analysis presented in Section 4 reveals this is not the case, and that with a certain iterate averaging scheme the population risk converges at the optimal rate. Consequently, it turns out the generalization gap is well bounded, and therefore that this variant constitutes a natural learning rule that is not stable in any sense but the most general one.

Without-replacement SGD for empirical risk minimization. The example featured in our main construction implies a lower bound of $\Omega(n^{-1/4})$ on the convergence rate of a single epoch of without-replacement SGD for finite sum optimization problems. In this setting, we have a set of n convex losses and we wish to minimize their sum by running SGD over random shufflings of the losses. While the smooth case has been studied extensively (e.g., [27, 26, 20, 30]), the non-smooth case has hardly received much attention. In Section 5 we extend our basic construction to a lower bound for the multi-epoch regime, and complement it with nearly matching upper bounds.

Our techniques. Fundamentally, we exploit the fact that dimension independent uniform convergence does not hold in SCO [31]. This is a prerequisite to any attempt at separating train and test losses of any hypothesis vector, let alone that produced by SGD. Another essential condition is the instability of SGD for non-smooth losses, as any form of stability would immediately imply a generalization gap upper bound regardless of uniform convergence. Our main lower bound draws inspiration from constructions presented in the works of [7] and [1], both of which rely on instability, the latter also exploiting failure of uniform convergence. However, neither of these contains the main ideas necessary to provoke the optimization dynamics required in our example. A crucial ingredient in our construction consists of encoding into the SGD iterate information about previous training examples. This, combined with careful design of the loss function, gradient oracle and population distribution, allows correlating sub-gradients of independent training examples, and in turn guiding the SGD iterates to *ascend* the empirical risk.

1.1 Summary of main contributions

To summarize, the main contributions of the paper are as follows:

- **One-pass SGD in SCO.** In Section 3, we study the basic SCO setup where the component losses are assumed to be individually convex, and present a construction where the expected empirical risk and therefore the generalization gap are both $\Omega(\eta\sqrt{n})$. We also provide extensions of our main construction demonstrating;
  - SCO with non-convex component functions may exhibit cases of benign *overfitting*, where $\mathbb{E}\left[F(\widehat{w}) - \widehat{F}(\widehat{w})\right] = \Omega(\eta^2 n)$.
  - In SCO with λ-strongly convex losses the worst case generalization gap is $\Omega(1/\lambda\sqrt{n})$ for the standard step size choice.
- **With vs without replacement SGD in SCO.** In Section 4, we prove the variant of SGD where the training examples are processed via sampling *with*-replacement from the *training set* minimizes the population risk at the optimal rate, and thus enjoys a generalization gap upper bound bound of $O(1/\sqrt{n})$.
- **Multi-epoch without-replacement SGD.** In Section 5, we study convergence rates of without-replacement SGD for finite sum convex optimization problems. We prove a lower bound of $\Omega(n^{-1/4}K^{-3/4})$ on the optimization error after K epochs over n convex losses, and complement with upper bounds of $O(n^{-1/4}K^{-1/2})$ and $O(n^{-1/4}K^{-1/4})$ for respectively the multi-shuffle and single-shuffle SGD variants.

1.2 Additional related work

Gradient descent, algorithmic stability and generalization. Closely related to our work is the study of stability properties of SGD. For smooth losses, [14] provide upper bounds on the generalization gap by appealing to uniform stability, yielding an $O(1/\sqrt{n})$ rate for a single epoch of n convex losses and the standard step size choice. In a later work, [7] prove tight rates for uniform stability of SGD in the setting of *non*-smooth losses, establishing these scale substantially worse; $\Theta(\eta\sqrt{n})$ for step size η and n training examples. Our work shows that in fact the worst case rate of the generalization gap completely coincides with the uniform stability rate of SGD.

A number of works prior to ours studied the extent to which SGD can be explained by implicit regularization in SCO. [16] study the setup where losses are smooth but only required to be convex in expectation, and show SGD may successfully learn when regularized ERM does not. Prior to

their work, [11] also rule out a wide range of implicit regularization based explanations of SGD in the basic SCO setup with convex losses. On a more general level, our work is related to the study of stability and generalization in modern learning theory, pioneered by [9, 31]. In particular, the failure of (dimension independent) uniform convergence in SCO was established in [31]. The work of [13] improves the dimension dependence in the construction of [31] from exponential to linear in the number of training examples. Notably, the construction featured in our main result requires the dimension to be exponential in the sample size, however the techniques of [13] do not readily extend to our setting. Thus, the optimal dimension dependence for a generalization gap lower bound is left for future work.

Without-replacement SGD for empirical risk minimization. A relatively long line of work studies convergence properties of without-replacement SGD from a pure optimization perspective (e.g., [27, 20, 29, 26, 19, 30]). Nearly all the papers in this line of work adopt the smoothness assumption, with near optimal bounds established by [20]. An exception is the paper of [32] where an $O(1/\sqrt{nK})$ upper bound is obtained for n datapoints and K epochs, albeit only for generalized linear models over a bounded domain — notably, a setting where uniform convergence holds. Prior to this thread of research, [22] prove a convergence rate of $O(n/\sqrt{K})$ for non-smooth loss functions that applies for *any* ordering of the losses. To the best of our knowledge, this is also the state-of-the-art result for without-replacement SGD in the non-smooth setting without further assumptions on the loss functions.

Benign overfitting vs. benign underfitting. While both benign underfitting and benign overfitting challenge traditional generalization techniques, that postulate the training error to represent the test error, as we discuss above these two phenomena point to very different regimes of learning. In particular, [33] shows that benign overfitting requires distributional assumptions for the interpolating algorithm to succeed. In contrast, we show that benign underfitting happens for SGD in a setting where it provably learns (namely, SCO), without any distributional assumptions. We also point out that Corollary 1 shows benign overfitting *cannot* happen in the setup we consider, hence the two phenomena seem to rise in different setups.

Explaining generalization of interpolators. As already discussed, there is a large recent body of work dedicated to understanding why over-parameterized models trained by SGD to zero training error generalize well [6, 8, and references therein]. In particular, the work of [5] aims at explaining the phenomenon for high dimensional linear models. Some recent papers investigate limitations of certain techniques in explaining generalization of interpolating algorithms: [21] show uniform convergence fails to explain generalization of SGD in a setup where the generalization gap is in fact well bounded, thus in sharp contrast to our work; [3] rule out the possibility of a large class of excess risk bounds to explain generalization of minimum norm interpolants. Unlike our work, they study properties of possible risk bounds when benign overfitting occurs, and thus do not pertain to SGD that never benignly overfits in SCO.

2 Preliminaries

We consider stochastic convex optimization (SCO) specified by a population distribution $\mathcal{Z}$ over a datapoint set Z, and loss function $f\colon W \times Z \to \mathbb{R}$ where $W \subset \mathbb{R}^d$ is convex and compact. We denote

$$F(w) \coloneqq \mathbb{E}_{z\sim\mathcal{Z}} f(w; z), \qquad \text{(population loss)}$$

$$\widehat{F}(w) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f(w; z_i), \qquad \text{(empirical loss)}$$

where $\{z_1, \dots, z_n\} \subseteq Z$ stands for the training set, which we regularly denote by S. We let $w^\star \coloneqq \min_{w\in W} F(w)$ denote the population minimizer, and $w_S^\star \coloneqq \min_{w\in W} \widehat{F}(w)$ denote the empirical risk minimizer (ERM). The diameter of W is defined by $\max_{x,y\in W}\{\|x-y\|\}$ where $\|\cdot\|$ denotes the euclidean norm, and $\mathcal{B}_0^d(1) \coloneqq \left\{x \in \mathbb{R}^d \mid \|x\| \le 1\right\}$ denotes the L_2 unit ball in $\mathbb{R}^d$. Given a training set $S = \{z_1, \dots, z_n\} \sim \mathcal{Z}^n$ and a learning algorithm that outputs a hypothesis $\widehat{w}_S$, we define the generalization gap to be the absolute value of the expected difference between test and train losses;

$$\left|\mathbb{E}_{S\sim\mathcal{Z}^n}\left[F(\widehat{w}_S) - \widehat{F}(\widehat{w}_S)\right]\right|. \qquad \text{(generalization gap)}$$

Throughout most of the paper, we consider one-pass projected SGD over S;

$$\text{initialize at } w_1 \in W;$$
$$\text{for } t = 2, \ldots, n: \quad w_{t+1} \leftarrow \Pi_W\left(w_t - \eta g_t\right), \quad \text{with } g_t \in \partial f(w_t; z_t),$$

where $\partial f(w; z)$ denotes the set of sub-gradients of $f(\cdot; z) \to \mathbb{R}$ at the point $w \in W$, and $\Pi_W : \mathbb{R}^d \to W$ the projection operation onto W.

3 A generalization gap lower bound for SGD

In this section, we establish our main result; that there exist convex learning problems where SGD incurs a large optimization error and therefore also a large generalization gap. When losses are convex these two quantities are closely related since in expectation, the empirical risk minimizer cannot significantly outperform the population minimizer (a claim that will be made rigorous shortly after our main theorem). Our construction builds on losses that are highly non-smooth, leading to SGD taking gradient steps that actually *ascend* the empirical objective.

Theorem 1. *Let $n \in \mathbb{N}$, $n \geq 4$, $d \geq 2^{4n \log n}$, and $W = \mathcal{B}_0^{2d}(1)$. Then there exists a distribution over instance set Z and a 4-Lipschitz convex loss function $f : W \times Z \to \mathbb{R}$ such that running SGD with step size $\eta > 0$ over $S \sim \mathcal{Z}^n$ yields;*

(i) a large optimization error; $\mathbb{E}\left[\widehat{F}(\widehat{w}_S) - \widehat{F}(w_S^\star)\right] = \Omega\left(\min\left\{\eta\sqrt{n}, \frac{1}{\eta\sqrt{n}}\right\}\right),$

(ii) a large generalization gap; $\mathbb{E}\left[\widehat{F}(\widehat{w}_S) - F(\widehat{w}_S)\right] = \Omega\left(\min\left\{\eta\sqrt{n}, \frac{1}{\eta\sqrt{n}}\right\}\right),$

where $\widehat{w}_S$ is any suffix average of the iterates. In particular, for $\eta = \Theta(1/\sqrt{n})$, the population risk is $\mathbb{E}\left[F(\widehat{w}_S) - F(w^\star)\right] = O(1/\sqrt{n})$, while the generalization gap and training error are both $\Omega(1)$.

A detailed proof of Theorem 1 is deferred to Appendix B; in the following we provide an informal overview containing its principal ingredients.

Proof sketch. Let $Z := \{0,1\}^d$, and consider a population distribution $\mathcal{Z}$ such that $z(i) = 1$ with probability δ. We will use a loss function of the form

$$f(w; z) := \|z \odot w\| + \phi(w; z),$$

where $\odot$ denotes element-wise product. The high level idea is that the norm component penalizes w's that correlate with the given sample point z, and the ϕ function (the details of which are left for Appendix B) is tailored so that it drives the SGD iterates precisely to those areas in the L_2 ball where it correlates with the training set $\{z_1, \ldots, z_n\}$. In addition, the choice of parameters is such that the population loss is approximately zero over the entire domain.

Taking d sufficiently large compared to δ^{-1}, we ensure that w.h.p., for every round $t \in [n]$ there exist many coordinates $i \in [d]$ with a prefix of ones; $z_1(i) = \cdots = z_{t-1}(i) = 1$. With δ chosen sufficiently small compared to n, we ensure that as long as $i \in [d]$ is any coordinate chosen independently of $\{z_{t+1}, \ldots, z_n\}$, w.h.p. this coordinate will have a suffix of zeros; $z_{t+1}(i) = \cdots = z_n(i) = 0$.

Our goal is to make SGD take steps $w_{t+1} \approx w_t - \eta e_{i_t}$ (where e_i denotes the i'th standard basis vector) where $i_t \in [d]$ is a coordinate with the aforementioned property of having a prefix of ones followed by a suffix of zeros. Note that since these steps are taken *after* the prefix of ones has ended, they will inflict large empirical loss from the norm component, but will not be "corrected" by future steps owed to the suffix of zeros. To achieve this, we design ϕ so that it encodes the relevant information into the SGD iterates. Specifically, ϕ "flags" (using some extra dimensions) all coordinates $i \in [d]$ where a prefix of ones has been encountered. In addition, using another max component in ϕ we have that for all such coordinates i, $e_i \in \partial f(w_t; z)$ for any example z (as this component in the loss depends only on the iterate w_t). In particular, we get that $e_i \in \partial f(w_t; z_t)$. Then, our gradient oracle just returns a subgradient pointing towards one of these coordinates (for convenience, we use the minimal one) which we denote by i_t, and SGD makes the desired step.

Notably, the coordinate i_t chosen by the subgradient oracle is independent of future examples, and therefore will have a suffix of zeros w.h.p. Hence, as mentioned, this ensures no gradient signal after round t will be able to correct the empirical risk ascent on i_t. Concluding, we have that for the final

iterate $\widehat{w} := w_{n+1}$, we get $\widehat{w}(i_t) = -\eta$ for all $t \in [n]$, therefore

$$\widehat{F}(\widehat{w}) = \frac{1}{n}\sum_{i=1}^{n} f(\widehat{w}; z_i) \approx \frac{1}{n}\sum_{i=1}^{n} \|z_i \odot \widehat{w}\| \approx \|\widehat{w}\| \approx \sqrt{\eta^2 n} = \eta\sqrt{n}.$$

A similar argument requiring a few more technical steps shows the same is true for any suffix average $\widehat{w}$. Noting that $\widehat{F}(0) = 0$, we get that the optimization error is $\Omega(\eta\sqrt{n})$. The implication for the generalization gap follows immediately with the standard step size choice of $\eta = 1/\sqrt{n}$, owed to SGD's population risk convergence guarantee. For an arbitrary step size, the result follows from a simple computation, and the proof is concluded. □

The magnitude of the generalization gap featured in Theorem 1 stems from the large optimization error, which results in the empirical risk over-estimating the population risk by a large margin. Evidently, for convex losses the converse is always false; the empirical risk will never significantly under-estimate the population risk (a fact that will turn out false when losses are only required to be convex in expectation — see Section 3.1). Indeed, stability of the regularized ERM solution implies the ERM does not perform significantly better on the training set compared to the population minimizer $w^\star$.

Lemma 1. *Let $W \subset \mathbb{R}^d$ with diameter D, $\mathcal{Z}$ any distribution over Z, and $f\colon W \times Z \to \mathbb{R}$ convex and G-Lipschitz in the first argument. Then $\mathbb{E}\left[\widehat{F}(w^\star) - \widehat{F}(w_S^\star)\right] \le \frac{4GD}{\sqrt{n}}$.*

Proof. Denote the regularized ERM by $\widehat{w}_S^\lambda := \arg\min_{w\in W}\left\{\frac{1}{n}\sum_{i=1}^n f_i(w; z_i) + \frac{\lambda}{2}\|w\|^2\right\}$. Observe,

$$F(w^\star) \le \mathbb{E}F(\widehat{w}_S^\lambda) \le \mathbb{E}\widehat{F}(\widehat{w}_S^\lambda) + \frac{4G^2}{\lambda n} \le \mathbb{E}\widehat{F}(w_S^\star) + \frac{\lambda}{2}D^2 + \frac{4G^2}{\lambda n},$$

where the second inequality follows from stability of the regularized ERM (see Lemma 13). Choosing $\lambda := 2GD/\sqrt{n}$, we get that

$$\mathbb{E}\left[\widehat{F}(w^\star) - \widehat{F}(w_S^\star)\right] = F(w^\star) - \mathbb{E}\widehat{F}(w_S^\star) \le \frac{4GD}{\sqrt{n}},$$

as claimed. □

Since the optimization error is always positive, we see that the upper bound given by Lemma 1 implies an upper bound on the difference between the population and empirical risks.

Corollary 1. *For any distribution $\mathcal{Z}$ over Z and Lipschitz loss function $f\colon W \times Z \to \mathbb{R}$ convex in the first argument, running SGD with step size $\eta := 1/\sqrt{n}$ guarantees $\mathbb{E}\left[F(\widehat{w}_S) - \widehat{F}(\widehat{w}_S)\right] \le O(1/\sqrt{n})$.*

Proof. We have,

$$\mathbb{E}\big[F(\widehat{w}_S) - \widehat{F}(\widehat{w}_S)\big] = \mathbb{E}\left[F(\widehat{w}_S) - F(w^\star)\right] + \mathbb{E}\big[\widehat{F}(w^\star) - \widehat{F}(\widehat{w}_S)\big]$$

The population error term on the RHS is $O(1/\sqrt{n})$ by the classical analysis of SGD. The second term is bounded by Lemma 1;

$$\mathbb{E}\big[\widehat{F}(w^\star) - \widehat{F}(\widehat{w}_S)\big] \le \mathbb{E}\big[\widehat{F}(w^\star) - \widehat{F}(w_S^\star)\big] \le 4GD/\sqrt{n},$$

and the result follows. □

In the subsections that follow we continue to study the generalization gap in the context of common variants to the basic SCO setup.

3.1 SCO with non-convex components

When we relax the convexity assumption and only require the losses to be convex in expectation, we can construct a learning problem where SGD exhibits a case of benign *overfitting*. In contrast to Theorem 1, here we actually drive the SGD iterates *towards* an ERM solution, thus achieving a low optimization error and an empirical risk that under-estimates the population risk.

Theorem 2. *Let $n \in \mathbb{N}$, $n \ge 4$, $d \ge 2^{4n\log n}$, $W = \mathcal{B}_0^{2d}(1)$, and $\eta \le 1/\sqrt{n}$. Then there exists a distribution $\mathcal{Z}$ over Z and a 4-Lipschitz loss $f\colon W \times Z \to \mathbb{R}$ where $\mathbb{E}_{z\sim\mathcal{Z}} f(w; z)$ is convex in w, such that for any suffix average $\widehat{w}$ of SGD with step size η;*

$$\mathbb{E}\left[F(\widehat{w}_S) - \widehat{F}(\widehat{w}_S)\right] = \Omega(\eta^2 n).$$

The construction and proof of Theorem 2 given in Appendix B.2 follow a methodology similar to that of Theorem 1. Here however, we exploit non convex losses to form an empirical loss landscape where the ERM solution significantly outperforms the population minimizer $w^\star$ (notably, a feat not possible when losses are individually convex, by Corollary 1). Our loss function is defined by $f(w;z) := \sum_{i=1}^d z(i)w(i)^2 + \phi(w;z)$, with each component playing a similar role as before. We work with the distribution $z \sim \{0,1\}^d$ where $z(i) = 1$ w.p. δ, $z(i) = -1$ w.p. δ, and $z(i) = 0$ w.p. $1-2\delta$. The intuition is that coordinates accumulating many -1's offer regions in the L_2 ball where the empirical risk is "too good" compared to the population risk. We tailor the extra dimensions and ϕ in coordination with the -1 values so that the sub-gradients guide the SGD iterates towards these regions, in exactly the same manner the construction of Theorem 1 drives the iterates to high loss regions. We note that while the statement of Theorem 2 is specialized to step size smaller than $1/\sqrt{n}$, it may be extended to any step size using arguments similar to those given in the proof of Theorem 1.

3.2 SCO with strongly convex components

Our basic construction extends to the strongly convex case by making only technical modification to Theorem 1. The theorem below concerns the standard step size choice for strongly convex objectives. We provide its proof in Appendix B.3.

Theorem 3. *Let $n \in \mathbb{N}$, $n \geq 10$, $d \geq 2^{4n\log n}$, $W = \mathcal{B}_0^{2d}(1)$, and $\lambda \geq 1/\sqrt{n}$. Then there exists a distribution over instance set Z and a 4-Lipschitz, λ-strongly convex loss function $f\colon W \times Z \to \mathbb{R}$ such that*

(i) the optimization error is large; $\mathbb{E}_{S\sim\mathcal{Z}^n}\left[\widehat{F}(\widehat{w}_S) - \widehat{F}(w_S^\star)\right] = \Omega\left(\frac{1}{\lambda\sqrt{n}}\right)$,

(ii) the generalization gap is large; $\mathbb{E}_{S\sim\mathcal{Z}^n}\left[\widehat{F}(\widehat{w}_S) - F(\widehat{w}_S)\right] = \Omega\left(\frac{1}{\lambda\sqrt{n}}\right)$,

where $\widehat{w}_S$ is any suffix average of SGD with step size schedule $\eta_t = 1/\lambda t$. Furthermore, the problem instance where this occurs is precisely the λ regularized version of the example featured in Theorem 1.

We note that an immediate implication of the above theorem is that if we seek a generalization gap upper bound for a weakly convex problem by means of regularization (meaning, by running SGD on a regularized problem), we would have to take $\lambda \geq 1$ to guarantee a gap of $O(1/\sqrt{n})$. To see this, note that the generalization gap (of any hypothesis) of the regularized problem is the same as that of the original. On the other hand, taking $\lambda \geq 1$ will of course be detrimental to the population error guarantee. Hence, one cannot circumvent the generalization gap lower bound by regularization without compromising the population error.

We conclude this section with a note regarding stability rates of SGD in non-smooth SCO. Implicit in Theorem 1, is that average stability of SGD coincides with the tight uniform stability rate of $\Theta(\eta\sqrt{n})$ established by [7]. This is because Theorem 1 provides the $\Omega(\eta\sqrt{n})$ lower bound on the most general stability notion, which is precisely the generalization gap [31]. We refer the reader to Appendix A for a more elaborate discussion.

4 SGD with vs without replacement

In this section, we consider a different algorithm in the context of the basic SCO setup; SGD over examples drawn *with*-replacement from the *training set*. This is not to be confused with one-pass SGD discussed in Section 3, which corresponds to without-replacement SGD on the training set, or alternatively with-replacement SGD over the *population* distribution. Given a training set $S = \{z_1, \ldots, z_n\} \sim \mathcal{Z}^n$, we define with-replacement projected SGD initialized at $w_1 \in W$ by

$$w_{t+1} \leftarrow \Pi_W\left(w_t - \eta\widehat{g}_t\right), \quad \text{where } \widehat{g}_t \in \partial f(w_t;\widehat{z}_t) \text{ and } \widehat{z}_t \sim \mathrm{Unif}(S).$$

Perhaps surprisingly, this version of SGD does not overfit the training data; our theorem below establishes that with proper iterate averaging, the population risk converges at the optimal rate.

Theorem 4. *Let $W \subset \mathbb{R}^d$ with diameter D, $\mathcal{Z}$ be any distribution over Z, and $f : W \times Z \to \mathbb{R}$ be convex and G-Lipschitz in the first argument. Let $S \sim \mathcal{Z}^n$ be a training set of $n \in \mathbb{N}$ datapoints drawn i.i.d. from $\mathcal{Z}$, and consider running SGD over training examples sampled with-replacement, uniformly and independently from S. Then, for step size $\eta = \frac{D}{G\sqrt{n}}$ and $\overline{w} := \frac{2}{n+1}\sum_{t=1}^n \frac{n-t+1}{n} w_t$, the following upper bound holds;*

$$\mathbb{E}\left[F(\overline{w}) - F(w^\star)\right] \leq \frac{10GD}{\sqrt{n}}.$$

Proof. Fix a time-step $t \in [n]$, and observe that if we don't condition on S, we may view the random datapoint $\widehat{z}_t$ as a mixture between a fresh i.i.d. sample from the population and a uniformly distributed sample from the previously processed datapoints $\widehat{S}_{t-1} := \{\widehat{z}_1, \ldots, \widehat{z}_{t-1}\}$;

$$\widehat{z}_t \mid \widehat{S}_{t-1} = \begin{cases} z \sim \mathfrak{Z} & \text{w.p. } 1 - \frac{t-1}{n}, \\ z \sim \mathrm{Unif}(\widehat{S}_{t-1}) & \text{w.p. } \frac{t-1}{n}. \end{cases}$$

With this in mind, denote $\widehat{f}_t(w) := f(w; \widehat{z}_t)$, fix $\widehat{S}_{t-1}$ and observe:

$$\mathbb{E}_{\widehat{z}_t}\left[\widehat{f}_t(w_t) - \widehat{f}_t(w^\star) \mid \widehat{S}_{t-1}\right] = \left(1 - \frac{t-1}{n}\right)\mathbb{E}_{z\sim\mathfrak{Z}}\left[f(w_t; z) - f(w^\star; z)\right] + \frac{t-1}{n}\frac{1}{t-1}\sum_{i=1}^{t-1} \widehat{f}_i(w_t) - \widehat{f}_i(w^\star).$$

Rearranging and taking expectation with respect to $\widehat{S}_{t-1}$ we obtain

$$\begin{aligned}\left(1 - \frac{t-1}{n}\right)\mathbb{E}\left[f(w_t; z) - f(w^\star; z)\right] &= \mathbb{E}\left[\widehat{f}_t(w_t) - \widehat{f}_t(w^\star)\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{t-1} \widehat{f}_i(w^\star) - \widehat{f}_i(w_t)\right] \\ &\leq \mathbb{E}\left[\widehat{f}_t(w_t) - \widehat{f}_t(w^\star)\right] + \frac{4GD\sqrt{t}}{n}, \end{aligned} \tag{2}$$

where the inequality follows from Lemma 1. Now, by a direct computation we have $\sum_{t=1}^{n}\left(1 - \frac{t-1}{n}\right) = \frac{n+1}{2}$, which motivates setting $\overline{w} := \frac{2}{n+1}\sum_{t=1}^{n}\frac{n-t+1}{n}w_t$. By convexity of F, Eq. (2), and the standard regret analysis of gradient descent [e.g., 15] we now have

$$\begin{aligned}\mathbb{E}\left[F(\overline{w}) - F(w^\star)\right] &\leq \frac{2}{n+1}\sum_{t=1}^{n}\left(1 - \frac{t-1}{n}\right)\mathbb{E}\left[F(w_t) - F(w^\star)\right] \\ &\leq \frac{2}{n+1}\sum_{t=1}^{n}\mathbb{E}\left[\widehat{f}_t(w_t) - \widehat{f}_t(w^\star)\right] + \frac{2}{n+1}\sum_{t=1}^{n}\frac{4GD\sqrt{t}}{n} \\ &\leq \frac{2}{n}\mathbb{E}\left[\sum_{t=1}^{n}\widehat{f}_t(w_t) - \widehat{f}_t(w^\star)\right] + \frac{8GD}{\sqrt{n}} \\ &\leq \frac{2}{n}\left(\frac{D^2}{2\eta} + \frac{\eta G^2}{2}\right) + \frac{8GD}{\sqrt{n}} \\ &= \frac{10GD}{\sqrt{n}}, \end{aligned}$$

where the last inequality follows by our choice of $\eta = \frac{D}{G\sqrt{n}}$. □

Evidently, the averaging scheme dictated by Theorem 4 does little to hurt the empirical risk convergence guarantee, which follows from the standard analysis with little modifications (for completeness we provide a formal statement and proof in Appendix C). Combined with Lemma 1, this immediately implies a generalization gap upper bound for with-replacement SGD. Notably, this shows with-replacement SGD provides for an example of a (natural) algorithm in the SCO learning setup that is not even stable on-average, but nonetheless has a well bounded generalization gap. We refer the reader to the discussion in Appendix A for more details.

Corollary 2. *For any distribution $\mathfrak{Z}$ and loss function $f : W \times Z \to \mathbb{R}$ convex and Lipschitz in the first argument, running SGD with step size and averaging as specified in Theorem 4 ensures*

$$\left|\mathbb{E}\left[F(\overline{w}) - \widehat{F}(\overline{w})\right]\right| \leq O(1/\sqrt{n}).$$

Proof. We have;

$$\left|\mathbb{E}\left[F(\overline{w}) - \widehat{F}(\overline{w})\right]\right| \leq \left|\mathbb{E}\left[F(\overline{w}) - F(w^\star)\right]\right| + \left|\mathbb{E}\left[\widehat{F}(w^\star) - \widehat{F}(w^\star_S)\right]\right| + \left|\mathbb{E}\left[\widehat{F}(w^\star_S) - \widehat{F}(\overline{w})\right]\right|.$$

The first term is upper bounded by convergence of the population risk provided by Theorem 4, the second by Lemma 1, and the third by the standard analysis of SGD (see Appendix C). □

5 Multi-epoch SGD for empirical risk minimization

In this section, we forgo the existence of a population distribution and discuss convergence properties of without-replacement SGD (wor-SGD) for finite sum optimization problems. A relatively long line of work discussed in the introduction studies this problem in the *smooth* case. The work of [20] noted smoothness is a necessary assumption to obtain rates that are strictly better than the $O(1/\sqrt{nK})$ guaranteed by with-replacement SGD for n losses and K epochs, due to a lower bound that follows from the deterministic case (e.g., [10]). Here we establish that smoothness is in fact necessary to obtain rates that are not *strictly worse* than with-replacement SGD. We consider running multiple passes of wor-SGD to solve the finite sum optimization problem given by the objective

$$F(w) := \frac{1}{n}\sum_{t=1}^{n} f(w;t) \tag{3}$$

where $\{f(w;t)\}_{t=1}^{n}$ is a set of n convex, G-Lipschitz losses defined over a convex and compact domain $W \subseteq \mathbb{R}^d$. Throughout this section we let $w^\star := \min_{w\in W} F(w)$ denote the minimizer of the objective Eq. (3). In every epoch $k \in [K]$ we process the losses in the order specified by a permutation $\pi_k : [n] \leftrightarrow [n]$ sampled uniformly at random, either once in the beginning of the algorithm (single-shuffle), or at the onset of every epoch (multi-shuffle). Multi-epoch wor-SGD initialized at $w_1^1 \in W$ is specified by the following equations;

$$\begin{aligned} w_{t+1}^k &\leftarrow w_t^k - \eta g_t^k, \quad \text{where } g_t^k \in \partial f_t^k(w_t^k) \\ w_1^{k+1} &:= w_{n+1}^k, \end{aligned}$$

where we denote $f_t^k(w) := f(w;\pi_k(t))$. A near-immediate implication of Theorem 1 is that there exists a set of convex losses on which a single epoch of wor-SGD cannot converge at a rate faster than $1/n^{1/4}$. Theorem 5 presented below extends our basic construction from Theorem 1 to accommodate multiple epochs. The main challenge here is in devising a mechanism that will allow fresh bad gradient steps to take place on every new epoch.

Theorem 5. *Let $n, K \in \mathbb{N}$, $K \geq 4, n \geq 4$, $c := 4/(2^{1/K}-1)$, $d \geq 2^{6n\log(cnK)}$, and $W = \mathcal{B}_0^{d'}(1)$ where $d' = (nK+1)d$. Then there exists a set of n convex, 4-Lipschitz losses such that after K epochs of either multi-shuffle or single-shuffle SGD, it holds that*

$$\mathbb{E}\left[F(\widehat{w}) - F(w^*)\right] = \Omega\left(\min\left\{1, \eta\sqrt{\frac{n}{J}} + \frac{1}{\eta nK} + \eta\right\}\right),$$

where $\widehat{w}$ is any suffix average of the last J epochs. In particular, we obtain a bound of $\Omega\left(n^{-1/4}K^{-3/4}\right)$ for any suffix average and any choice of η.

The proof of Theorem 5 is provided in Appendix D.1. The construction in the proof takes the idea that the training set can be encoded in the SGD iterate to the extreme. The loss function and gradient oracle are designed in such a way so as to record the training examples in their full form and order into the iterate. We then exploit this encoded information with an "adversarial" gradient oracle that returns the bad sub-gradients on each gradient step in every new epoch.

Next, we complement Theorem 5 with an upper bound that builds on stability arguments similar to those of the smooth case [20]. Importantly though, lack of smoothness means worse stability rates and necessitates extra care in the technical arguments. Below, we prove the multi-shuffle case, and defer the full details for the single-shuffle case to Appendix D.2.

Theorem 6. *Let $S = \{f(w;t)\}_{t=1}^{n}$ be a set of n convex, G-Lipschitz losses over a convex and compact domain $W \subseteq \mathbb{R}^d$ of diameter D, and consider running $K \geq 1$ epochs of wor-SGD over S. Then, we have the following guarantees:*

(i) For multi-shuffle, with step-size $\eta = D/(Gn^{3/4}K^{1/2})$, we have

$$\mathbb{E}\left[F(\widehat{w}) - F(w^\star)\right] \leq \frac{3GD}{n^{1/4}K^{1/2}}.$$

(ii) For single-shuffle, with step-size $\eta = D/(2Gn^{3/4}K^{3/4})$ and assuming $K \geq n$, we have

$$\mathbb{E}\left[F(\widehat{w}) - F(w^\star)\right] \leq \frac{10GD}{n^{1/4}K^{1/4}}.$$

In both of the above bounds, $\widehat{w} = \frac{1}{nK}\sum_{k\in[K],t\in[n]} w_t^k$, and the expectation is over the random permutations of losses.

Proof (multi-shuffle case). Observe;

$$\begin{aligned}
F(\widehat{w}) - F(w^\star) &\leq \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} F(w_t^k) - F(w^\star) \\
&= \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} F(w_t^k) - f_t^k(w^\star) \\
&= \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} F(w_t^k) - f_t^k(w_t^k) + \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} f_t^k(w_t^k) - f_t^k(w^\star) \\
&\leq \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} F(w_t^k) - f_t^k(w_t^k) + \frac{D^2}{2\eta nK} + \frac{\eta G^2}{2},
\end{aligned}$$

with the last inequality following from the standard nK round regret bound for gradient descent [see e.g., 15]. To bound the other term, using Lemma 10, we relate the difference between the without-replacement loss distribution and the full batch objective to the uniform stability rate of SGD, which may then be bounded by applying Lemma 11:

$$\begin{aligned}
\mathbb{E}\left[F(w_t^k) - f_t^k(w_t^k))\right] &= \mathbb{E}_{\pi_1,\ldots,\pi_{k-1}}\mathbb{E}_{\pi_k}\left[F(w_t^k) - f_t^k(w_t^k) \mid w_1^k\right] \\
&\leq \mathbb{E}_{\pi_1,\ldots,\pi_{k-1}}\left[G\epsilon_{\text{stab}}^{\text{SGD}}(t-1)\right] \\
&= G\epsilon_{\text{stab}}^{\text{SGD}}(t-1) \\
&\leq 2\eta G^2\sqrt{t}.
\end{aligned}$$

Concluding, we have that

$$\begin{aligned}
\mathbb{E}\left[F(\widehat{w}) - F(w^\star)\right] &\leq \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n}\mathbb{E}\left[F(w_t^k) - f_t^k(w_t^k)\right] + \frac{D^2}{2\eta nK} + \frac{\eta G^2}{2} \\
&\leq \frac{2}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n}\eta G^2\sqrt{t} + \frac{D^2}{2\eta nK} + \frac{\eta G^2}{2} \\
&\leq 2\eta G^2\sqrt{n} + \frac{D^2}{2\eta nK} + \frac{\eta G^2}{2} \\
&\leq \frac{3GD}{n^{1/4}K^{1/2}},
\end{aligned}$$

where the last inequality follows from our choice of $\eta = D/(Gn^{3/4}K^{1/2})$.

□

Acknowledgements and funding disclosure

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grants number 993/17, 2549/19, 2188/20), by the Len Blavatnik and the Blavatnik Family foundation, by the Yandex Initiative in Machine Learning at Tel Aviv University, and by an unrestricted gift from Google. Any opinions, findings, and conclusions or recommendations expressed in this work are those of the author(s) and do not necessarily reflect the views of Google.

References

[1] I. Amir, T. Koren, and R. Livni. SGD generalizes better than GD (and regularization doesn't help). In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 63–92. PMLR, 2021.

[2] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

[3] P. L. Bartlett and P. M. Long. Failures of model-dependent generalization bounds for least-norm interpolation. *Journal of Machine Learning Research*, 22(204):1–15, 2021.

[4] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[5] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[6] P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021.

[7] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

[8] M. Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv preprint arXiv:2105.14368*, 2021.

[9] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

[11] A. Dauber, M. Feder, T. Koren, and R. Livni. Can implicit bias explain generalization? stochastic convex optimization as a case study. *Advances in Neural Information Processing Systems*, 33, 2020.

[12] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017*, 2018.

[13] V. Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[14] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[15] E. Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

[16] S. Kale, A. Sekhari, and K. Sridharan. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *arXiv preprint arXiv:2107.05074*, 2021.

[17] N. Littlestone and M. Warmuth. Relating data compression and learnability, 1986.

[18] D. A. McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

[19] K. Mishchenko, A. Khaled Ragab Bayoumi, and P. Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] D. Nagaraj, P. Jain, and P. Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711. PMLR, 2019.

[21] V. Nagarajan and J. Z. Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] A. Nedić and D. Bertsekas. Convergence rate of incremental subgradient algorithms. In *Stochastic optimization: algorithms and applications*, pages 223–264. Springer, 2001.

[23] A. S. Nemirovskij and D. B. Yudin. Problem complexity and method efficiency in optimization, 1983.

[24] G. Neu. Information-theoretic generalization bounds for stochastic gradient descent. *arXiv preprint arXiv:2102.00931*, 2021.

[25] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[26] S. Rajput, A. Gupta, and D. Papailiopoulos. Closing the convergence gap of SGD without replacement. In *International Conference on Machine Learning*, pages 7964–7973. PMLR, 2020.

[27] B. Recht and C. Ré. Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. In *Conference on Learning Theory*, pages 11–1. JMLR Workshop and Conference Proceedings, 2012.

[28] D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240. PMLR, 2016.

[29] I. Safran and O. Shamir. How good is SGD with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.

[30] I. Safran and O. Shamir. Random shuffling beats sgd only after many epochs on ill-conditioned problems. *arXiv preprint arXiv:2106.06880*, 2021.

[31] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[32] O. Shamir. Without-replacement sampling for stochastic gradient methods. *Advances in neural information processing systems*, 29:46–54, 2016.

[33] O. Shamir. The implicit bias of benign overfitting. *arXiv preprint arXiv:2201.11489*, 2022.

[34] U. Sherman, T. Koren, and Y. Mansour. Optimal rates for random order online optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[35] V. Vapnik. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–281, 1971.

[36] A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2521–2530, 2017.

[37] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [N/A]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Relations to Algorithmic Stability of SGD

In this section, we formally introduce notions of algorithmic stability and relate them to results presented in the paper. Let Z denote a set of datapoints and $\mathcal{Z}$ a distribution over Z. For two training sets $S, S' \in Z^n$, we write $S \simeq S'$ if they differ in exactly one datapoint. For a learning algorithm $\mathcal{A}\colon Z^* \to \mathbb{R}^d$, we define the *uniform argument stability* (UAS) of $\mathcal{A}$ by

$$\epsilon_{\text{stab}}^{\mathcal{A}}(n) := \max_{S \simeq S', |S|=n} \left\|\mathcal{A}(S) - \mathcal{A}(S')\right\|, \tag{4}$$

and the *average argument stability* (AAS) of $\mathcal{A}$ by

$$\epsilon_{\text{avgstab}}^{\mathcal{A}}(n) := \max_{\mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{S\sim\mathcal{Z}^n, z_i'\sim\mathcal{Z}} \left\|\mathcal{A}(S) - \mathcal{A}(S^{(i)})\right\| \right\}, \tag{5}$$

where $S^{(i)}$ is formed by taking S and replacing z_i with z_i'.

It is well known (e.g., [9, 31]) that for any distribution $\mathcal{Z}$ and algorithm $\mathcal{A}$, the following relations holds between the generalization gap, AAS, and UAS;

$$\left|\mathbb{E}_{S\sim\mathcal{Z}^n}\left[F(\mathcal{A}(S)) - \widehat{F}(\mathcal{A}(S))\right]\right| \le \epsilon_{\text{avgstab}}^{\mathcal{A}}(n) \le \epsilon_{\text{stab}}^{\mathcal{A}}(n). \tag{6}$$

In [7] it was established the UAS of both with and without-replacement SGD is $\Omega(\eta\sqrt{n})$ for n steps of size η with n training examples. However, considering Eq. (6), it remained unclear whether the AAS and generalization gap of these algorithms exhibit rates of similar order, in which case the UAS accurately captures the rate of the generalization gap. Interestingly, the answer to this question depends on whether sampling is done with or without replacement, as we discuss next.

Stability of without-replacement SGD. As an immediate corollary of our Theorem 1, we have that the AAS of without-replacement SGD is also $\Omega(\eta\sqrt{n})$. This follows from Eq. (6) and since the theorem establishes the generalization gap to be $\Omega(\eta\sqrt{n})$. Similarly, the lower bound given by Theorem 3 demonstrates that the AAS of SGD in the strongly convex case is $\Omega(1/\lambda\sqrt{n})$. Combined with the naive upper bound argument for uniform stability of $O(1/\lambda\sqrt{n})$ (which follows by convergence of SGD iterates to the minimizer in parameter space), we get a tight characterization of stability for strongly convex losses for the standard step size schedule.

Stability of with-replacement (one-pass) SGD. In Section 4, specifically in Corollary 2, we establish a generalization gap of $O(1/\sqrt{n})$ for with-replacement SGD with a particular averaging scheme and a properly tuned step size. However, as it turns out, the average argument stability of this version of SGD is nonetheless of order $\Omega(\eta\sqrt{n})$ as we demonstrate in Theorem 7 below. This shows that this version of with-replacement SGD is an algorithm that is not stable in any sense but the most general one (namely, the one being equivalent to the generalization gap).

Theorem 7. *Let $n \in \mathbb{N}$, $n \ge 200$, $d \ge 2^{3n}$, $W = \mathcal{B}_0^d(1)$. Further, let $\{\beta_t\}_{t=1}^n$ be an iterate averaging scheme that does not decay too quickly; $\sum_{s=t+1}^n \beta_n \ge C((n-t)/n)^2$ for some constant $C > 0$. Then there exists a distribution over the instance set $Z = \{0,1\}^d$ and a 2-Lipschitz, convex loss function $f\colon W \times Z \to \mathbb{R}$ such that for all $k \in [n]$,*

$$\mathbb{E}_{S\sim\mathcal{Z}^n, z_k'\sim\mathcal{Z}}\left\|\widehat{w} - \widehat{w}^{(k)}\right\| \ge \Omega(\eta\sqrt{n}),$$

where $\widehat{w}, \widehat{w}^{(k)}$ denote the $\{\beta_n\}$-averaged iterates $\{w_t\}, \{w_t'\}$ of n with-replacement SGD iterations (initialized at $w_1' = w_1 = 0$) with step size $\eta > 0$ over the training sets S and $S^{(k)}$ respectively;

$$\widehat{w} := \sum_{t=1}^{n} \beta_n w_t; \quad \widehat{w}^{(k)} := \sum_{t=1}^{n} \beta_n w_t'.$$

Note that the averaging scheme employed in Theorem 4 decays sufficiently slow so as to satisfy requirements of Theorem 7, hence the stability lower bound follows.

Proof. Let $\mathcal{Z}$ be defined by $z(i) \sim \text{Ber}(1/2)$, and set

$$f(w; z) := -\epsilon \sum_{i=1}^{d} \alpha_z(i) w(i) + \max_{i\in[d]}\{w(i)\}.$$

We will take $\epsilon := \beta_n/(16n^3 d)$, and define

$$\alpha_z(i) = \begin{cases} -n & z(i) = 1, \\ 1 & z(i) = 0. \end{cases}$$

In addition, let

$$I(w) := \arg\min_{i \in [d]} \left\{ i \mid w(i) = \max_j \{w(j)\} \right\},$$

set

$$i_t := I(w_t), \quad i'_t := I(w'_t), \tag{7}$$

and define

$$g(w; z) := -\epsilon \alpha_z + e_{I(w)}, \tag{8}$$

where e_i denotes the i'th standard basis vector. It follows that $g(w; z) \in \partial f(w; z)$ and $\|g(w; z)\| \leq nd\epsilon + 1 \leq 2$ for all $z \in Z, w \in W$. Proceeding, we denote

$$S = \{z_1, \ldots, z_n\}, \quad S' = \{z'_1, \ldots, z'_n\},$$

and note that $z_l = z'_l$ for all $l \neq k$. Further, we denote the training examples sampled by SGD by

$$\widehat{z}_t := z_{k_t} \in S, \quad \widehat{z}'_t := z'_{k_t} \in S^{(k)},$$

where $\{k_t\} \sim \text{Unif}[n]$ are uniformly random and independent training indices. For the remainder of the proof, we condition on the event

$$\mathcal{E} = \{(z_1, \ldots, z_n, z'_k) \mid Z(S) \geq n, \text{ and } Z(S') \geq n\}, \tag{9}$$

where $Z(S) := |\{i \mid \forall r \in [n],\ z_r(i) = 0\}|$ and similarly $Z(S') := |\{i \mid \forall r \in [n],\ z'_r(i) = 0\}|$. Owed to our assumption that $d \geq 2^{3n}$, a standard concentration argument shows this event occurs with probability $\geq 1/2$.

We will now proceed to track how the SGD iterates evolve. Observe that for all $t \in [n]$, we have by direct computations of the gradient steps with Eq. (8);

$$\begin{aligned} i \notin \{i_1, \ldots, i_t\} \implies w_{t+1}(i) &= \eta\epsilon \sum_{s=1}^{t} \alpha_{\widehat{z}_s}(i) \\ &\implies \begin{cases} w_{t+1}(i) = \eta\epsilon t & \forall s \leq t,\ \widehat{z}_s(i) = 0, \\ w_{t+1}(i) \in [-\eta\epsilon n^2, 0] & \exists s \leq t,\ \widehat{z}_s(i) = 1. \end{cases} \end{aligned}$$

In addition, from similar computations;

$$i \in \{i_1, \ldots, i_t\} \implies w_{t+1}(i) \leq -\eta + \eta\epsilon n.$$

Summarizing, and applying identical calculations for w'_{t+1}, we have:

$$\begin{aligned} t < s \implies w_s(i_t) \leq -\eta + \eta\epsilon n, \quad i \notin \{i_1, \ldots, i_n\} &\implies \forall s,\ w_s(i) \in [-\eta\epsilon n^2, \eta\epsilon n] \\ t < s \implies w'_s(i'_t) \leq -\eta + \eta\epsilon n, \quad i' \notin \{i'_1, \ldots, i'_n\} &\implies \forall s,\ w'_s(i') \in [-\eta\epsilon n^2, \eta\epsilon n]. \end{aligned} \tag{10}$$

By Eq. (10) above, for all $t \in [n]$ we have;

$$\begin{aligned} \widehat{w}(i_t) = \sum_{s=1}^{n} \beta_s w_s(i_t) &\leq \sum_{s=1}^{t} \beta_s \eta s\epsilon - \sum_{s=t+1}^{n} \beta_s(\eta - \eta\epsilon n) \\ &\leq -\eta \sum_{s=t+1}^{n} \beta_s + \eta n\epsilon \sum_{s=1}^{n} \beta_s \leq -\frac{3}{4}\eta \sum_{s=t+1}^{n} \beta_s, \end{aligned} \tag{11}$$

where the last inequality follows from our choice of ϵ. In addition, if $i_t \notin \{i'_1, \ldots i'_n\}$, again by Eq. (10) and our choice of ϵ it follows that;

$$\widehat{w}^{(k)}(i_t) = \sum_{s=1}^{n} \beta_s w'_s(i_t) \geq \sum_{s=1}^{n} \beta_s \eta n^2 \epsilon \geq -\frac{\eta}{4}\beta_n. \tag{12}$$

Now, set $t_0 \coloneqq \min\{t \mid k_t = k\}$ to be the first time that index k (in which training examples differ) is chosen, and let $t > t_0$. Note that $z'_k(i_t) = 1$ implies $i_t \notin \{i'_1, \ldots, i'_n\}$; to see this, observe that for $\tau \le t_0$, $i'_\tau = i_\tau$, while $\tau > t_0$ implies $i'_\tau \neq i_t$, since the event Eq. (9) we condition on ensures

$$i_t = \min\{i \in [d] \mid \forall s < t,\ \widehat{z}_s(i) = 0\},$$
$$i'_t = \min\{i \in [d] \mid \forall s < t,\ \widehat{z}'_s(i) = 0\}.$$

(From the above it also must hold that $i_t \neq i_\tau, i'_t \neq i'_\tau$ for all $t \neq \tau$.) Thus, putting together Eq. (11), Eq. (12) and the fact that $z'_k(i_t) = 1$ implies $i_t \notin \{i'_1, \ldots, i'_n\}$, we obtain for all $t_0 < t$;

$$|\widehat{w}(i_t) - \widehat{w}^{(k)}(i_t)| \ge \mathbb{1}\left\{z'_k(i_t) = 1\right\} \frac{\eta}{2} \sum_{s=t+1}^{n} \beta_s \ge \mathbb{1}\left\{z'_k(i_t) = 1\right\} \frac{\eta C (n-t)^2}{2n^2},$$

where the second inequality follows from our assumption on $\{\beta_t\}$. Thus, for $t_0 < t \le 3n/4$, we get that

$$|\widehat{w}(i_t) - \widehat{w}^{(k)}(i_t)| \ge \mathbb{1}\left\{z'_k(i_t) = 1\right\} \frac{C}{4}\eta,$$

and taking expectations we obtain;

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{w} - \widehat{w}^{(k)}\right\| \mid t_0\right] &\ge \mathbb{E}\left[\sqrt{\sum_{t=1}^{n} (\widehat{w}(i_t) - \widehat{w}'(i_t))^2} \mid t_0\right] \\
&\ge \frac{C\eta}{4} \mathbb{E}\left[\sqrt{\sum_{t=t_0+1}^{3n/4} \mathbb{1}\left\{z'_k(i_t) = 1\right\}} \mid t_0\right].
\end{aligned}$$

Now, observe that z'_k is independent of i_t for all t, hence the expectation above is of the form

$$\mathbb{E}\sqrt{\sum_{l=1}^{m} Y_l},$$

where $m \coloneqq 3n/4 - t_0$ and $Y_l \sim \mathrm{Ber}(1/2)$ are independent. Thus,

$$\Pr\left(m/2 - \sum_{l=1}^{m} Y_t > m/4\right) \le e^{-m/16} \le 1/2,$$

for $m > 100$, and then

$$\mathbb{E}\left[\left\|\widehat{w} - \widehat{w}^{(k)}\right\| \mid t_0\right] \ge \frac{C\eta}{4}\frac{1}{2}\sqrt{3n/4 - t_0} = \frac{C\eta}{8}\sqrt{3n/4 - t_0}.$$

To conclude, we use the fact that t_0 follows a geometric distribution with parameter $1/n$, therefore

$$\Pr(t_0 \le n/2) = \frac{1}{n}\sum_{t=1}^{n/2}(1 - 1/n)^t = 1 - (1 - 1/n)^{n/2} \ge 1 - e^{-1/2} \ge 1/3.$$

This implies,

$$\mathbb{E}\left\|\widehat{w} - \widehat{w}^{(k)}\right\| \ge \frac{C\eta}{32}\sqrt{3n/4 - n/2} = \frac{C\eta}{64}\sqrt{n},$$

and completes the proof. □

B Proof of Theorem 1

Our first proof below applies for step sizes $\eta \le 1/\sqrt{n}$. The extension for larger step sizes is rather technical and requires care of the projection step — we provide it in Appendix B.1. The statement of Theorem 1 is repeated below for the case of the small step size regime.

Theorem 8 (Small step size case of Theorem 1)**.** *Let $n \in \mathbb{N}$, $n \geq 4$, $d \geq 2^{4n \log n}$, and $W = \mathcal{B}_0^{2d}(1)$. Then there exists a distribution over instance set Z and a 4-Lipschitz convex loss function $f : W \times Z \to \mathbb{R}$ such that*

(i) the optimization error is large; $\mathbb{E}_{S \sim \mathcal{Z}^n}\left[\widehat{F}(\widehat{w}_S) - \widehat{F}(w_S^\star)\right] = \Omega\left(\eta\sqrt{n}\right)$,

(ii) the generalization gap is large; $\mathbb{E}_{S \sim \mathcal{Z}^n}\left[\widehat{F}(\widehat{w}_S) - F(\widehat{w}_S)\right] = \Omega\left(\eta\sqrt{n}\right)$,

where $\widehat{w}$ is any suffix average of SGD with step size $\eta \leq 1/\sqrt{n}$.

Proof. Our construction is parameterized by $\epsilon, \delta > 0$, which will be chosen later. We will work with the datapoints set $Z = \{0,1\}^{2d}$ and define the distribution $\mathcal{Z} = \mathcal{Z}(\delta)$ over Z by

$$\forall i \leq d; \quad z(i) = \begin{cases} 1 & \text{w.p. } \delta; \\ 0 & \text{w.p. } (1-\delta), \end{cases}$$
$$\forall i > d; \quad z(i) = z(i-d).$$

Our loss function is a combination of two components; the "push" function ϕ is in charge of driving the SGD iterate towards areas in the L_2 ball where the "penalty" function ν inflicts a norm-like loss.

$$\phi(w; z) := -\epsilon \sum_{i=d+1}^{2d} z(i) w(i) + \max_{1 \leq i \leq d} \{w(i) + w(i+d)\}, \tag{13}$$

$$\nu_z(w) := \sqrt{\sum_{i=1}^{d} z(i) w(i)^2}, \tag{14}$$

$$f(w; z) := \phi(w; z) + \nu_z(w).$$

The lower bound arguments all go through with any sub-gradient oracle $g(w; z) \in \partial_w f(w; z)$. For clarity of exposition, we make use of the gradient oracle g that returns the minimal coordinate sub-gradient for the max component in ϕ;

$$g_\phi(w; z)(i) := \begin{cases} \mathbb{1}\{i = I(w)\} & i \leq d \\ -\epsilon z(i) + \mathbb{1}\{i = I(w) + d\} & i \geq d \end{cases}, \tag{15}$$

$$\text{where } I(w) := \min\left\{i \in [d] \mid i \in \operatorname*{arg\,max}_{1 \leq j \leq d} \{w(j) + w(j+d)\}\right\}. \tag{16}$$

We additionally denote the index picked by g on round $t \in [n]$ by

$$i_t := I(w_t). \tag{17}$$

We then set $g(w; z) := g_\phi(w; z) + \nabla \nu_z(w)$. It now follows that for all $w, z \in \mathbb{R}^{2d}$, $g_\phi(w; z) \in \partial_w \phi(w; z)$, thus $g(w; z) \in \partial_w f(w; z)$. Choosing $\epsilon = 1/d$, we get that f is 4-Lipschitz;

$$\|g(w; z)\| \leq \epsilon\sqrt{d} + 2 + \frac{1}{2\|w\|}\|w\| \leq 3 + 1/\sqrt{d} \leq 4.$$

With the above construction in place, we first claim that with sufficiently large probability, the training set will contain the desired collection of "bad" coordinates which will be picked up by our gradient oracle. Indeed, with the dimension d large enough, a proper choice of δ ensures that for every $t \in [n]$, there will be a certain coordinate with a prefix of $t-1$ ones followed by a zero only suffix.

Lemma 2. For $\delta = 1/4n^2$, with probability $\geq 1/2$ over the random draw of $S = \{z_1, \ldots, z_n\} \sim \mathcal{Z}^n$, it holds that for all $t \in [n]$:

1. *There exist prefix of ones coordinates $E^{(1,t)} := \{j \in [d] \mid s < t \implies z_s(j) = 1\} \neq \emptyset$, and*
2. *the minimal such coordinate $J_t := \min\left\{j \in E^{(1,t)}\right\}$ also has a zero suffix; $s \geq t \implies z_s(J_t) = 0$.*

From this point onward fix $\delta := 1/4n^2$. By definition of our gradient oracle, a relatively straightforward argument given in our next lemma establishes SGD will take gradient steps precisely on those bad coordinates of Lemma 2. Notably, we have designed the construction so that these steps are made only after the samples penalizing those coordinates have been processed. This eliminates the possibility for SGD to correct these coordinates in future steps.

Lemma 3. We have with probability $\geq 1/2$ that for all $t \in [n]$, $i_t = J_t$ (see Eq. 17), and for all $\tau \in [n], \tau > t$, $w_\tau(i_t) = -\eta$.

To complete the proof, we assume the event from the lemma occurs. Since it occurs with constant probability, a lower bound derived conditioned on it implies a lower bound in expectation. First we argue the population loss of all iterates is upper bounded as

$$\begin{aligned}
F(w_t) &= -\epsilon \sum_{i=d+1}^{2d} \delta w_t(i) + \epsilon n\eta + \mathbb{E}\left[\sqrt{\sum_{i=1}^{d} z(i) w_t(i)^2}\right] \\
&\leq -\epsilon\delta \sum_{s=1}^{t-1} w_t(i_s) + \epsilon n\eta + \sqrt{\sum_{i=1}^{d} \mathbb{E}[z(i)] w_t(i)^2} \\
&= \epsilon\delta(t-1)\eta + \epsilon n\eta + \sqrt{\delta \sum_{i=1}^{d} w_t(i)^2} \\
&\leq 2\epsilon n\eta + \sqrt{\delta \sum_{s=1}^{t} w_t(i_s)^2} \\
&\leq 2\epsilon n\eta + \sqrt{\delta\eta^2 n} = \frac{2n\eta}{d} + \frac{\eta\sqrt{n}}{2n} \leq \frac{\eta}{\sqrt{n}}.
\end{aligned}$$

By convexity of the population loss, the above implies that any suffix average satisfies $F(\widehat{w}) \leq \eta/\sqrt{n}$. In addition, note that $\widehat{F}(w_S^\star) \leq \widehat{F}(0) \leq 0$, hence the $\Omega(\eta\sqrt{n})$ bound we will now establish on the empirical risk of SGD implies our claimed optimization and generalization lower bounds. Indeed, let $\tau \in [n]$ and denote $\overline{w}_\tau := \frac{1}{n-\tau+2}\sum_{t=\tau}^{n+1} w_t$. Observe that for $1 \leq t \leq n/2$, by Lemma 3 at least half of the iterates have the $-\eta$ value in coordinate i_t;

$$\begin{aligned}
\overline{w}_\tau(i_t) = \frac{1}{n-\tau+2}\sum_{s=\tau}^{n+1} w_s(i_t) &\leq \frac{1}{n-\tau+2}\sum_{s=\max\{\tau, n/2\}}^{n+1} w_s(i_t) \\
&\leq \frac{n - \max\{\tau, n/2\} + 2}{n-\tau+2}(-\eta) \leq -\frac{\eta}{2}.
\end{aligned}$$

(We ignore the fact that the last iterate, formally speaking, may have a slightly greater value due to the projection on the last step.) Now, for any $w \in W$,

$$\begin{aligned}
\widehat{F}(w) = \frac{1}{n}\sum_{s=1}^{n} f(w; z_s) &= \frac{1}{n}\sum_{s=1}^{n} \phi(w; z_s) + \frac{1}{n}\sum_{s=1}^{n}\sqrt{\sum_{t=s+1}^{n} w(i_t)^2} \\
&\geq \frac{1}{n}\sum_{s=1}^{n} \phi(w; z_s) + \frac{1}{5\sqrt{n}}\sum_{t=n/4}^{n} |w(i_t)|,
\end{aligned}$$

where the second inequality follows from Lemma 15. Noting that $\phi(\overline{w}_\tau; z_s) \geq -\epsilon^2 dn\eta \geq -\eta$ and combining the last two displays we get that

$$\widehat{F}(\overline{w}_{\tau:n}) \geq \frac{1}{5\sqrt{n}}\sum_{t=n/4}^{n/2} \frac{\eta}{2} - \eta = \frac{\eta(\sqrt{n}-1)}{40},$$

which completes the proof. □

Proof of Lemma 2. Fix $t \in [n]$, denote

$$\begin{aligned}
E^{(1,t)} &:= \left\{i \in [d] \mid z_s(i) = 1 \ \forall s < t\right\}, \\
E^{(t,0)} &:= \left\{i \in [d] \mid z_s(i) = 0 \ \forall s \geq t\right\},
\end{aligned}$$

and let $J_t \in E^{(1,t)}$ be the minimal element if it is not empty. Note that

$$\Pr(E^{(1,t)} = \emptyset) = \Pr\left(\forall i \in [d],\ \exists s < t,\ z_s(i) = 0\right) = \left(1 - \delta^{t-1}\right)^d \leq (1-\delta^n)^d.$$

In addition, since the contents of $E^{(1,t)}$ are independent of $z_t, \ldots, z_n$, we have that for any $i \in E^{(1,t)}$,

$$\Pr(i \in E^{(t,0)}) = (1-\delta)^{n-t+1} \geq (1-\delta)^n.$$

Therefore,

$$\Pr\left(E^{(1,t)} = \emptyset \text{ OR } \left(E^{(1,t)} \neq \emptyset \text{ but } J_t \notin E^{(t,0)}\right)\right) \leq (1-\delta^n)^d + 1 - (1-\delta)^n.$$

Now, by the union bound over all values of $t \in [n]$ we obtain

$$\Pr\left(\forall t \in [n],\ E^{(1,t)} \neq \emptyset \text{ AND } J_t \in E^{(t,0)}\right) \geq 1 - n\left((1-\delta^n)^d + 1 - (1-\delta)^n\right). \tag{18}$$

Now, since $\delta = 1/4n^2$ we have

$$(1-\delta^n)^d = \left(1 - \frac{d\delta^n}{d}\right)^d \leq e^{-d\delta^n} \leq \frac{1}{4n},$$

where the last inequality follows for $d \geq \delta^{-n}\log(4n) = 4^n n^{2n}\log(4n)$ (recall that by the assumption in the theorem statement $d \geq 2^{4n\log n} \geq 4^n n^{2n} 4n$). In addition;

$$(1-\delta)^n = \left(1 - \frac{1}{4n^2}\right)^n \geq 1 - \frac{1}{4n} \implies 1 - (1-\delta)^n \leq \frac{1}{4n}.$$

Back to Eq. (18), applying the inequalities from the last two displays we obtain the desired event occurs with probability

$$\geq 1 - n\left((1-\delta^n)^d + 1 - (1-\delta)^n\right) \geq 1 - n\left(\frac{1}{4n} + \frac{1}{4n}\right) = \frac{1}{2}, \tag{19}$$

and the result follows. □

Proof of Lemma 3. Following a direct computation, we get that

$$g(w;z)(i) = \begin{cases} \mathbb{1}\{i = I(w)\} + \frac{z(i)w(i)}{v_z(w)} & i \leq d, \\ \mathbb{1}\{i = I(w) + d\} - \epsilon z(i) & i > d. \end{cases} \tag{20}$$

From the above we see that the value of $w(i+d)$ for every coordinate $i + d \in \{d+1, \ldots, 2d\}$ gains $\eta\epsilon$ when $z(i) = 1$, while the value of coordinate i only decreases. Thus $I(w_t)$ will be a coordinate with an all ones prefix if one exists. Formally, let $t \in [n]$, and observe that our gradient oracle will return the minimal coordinate $i_t \in [d]$ with the maximum value of $w_t(i_t) + w_t(i_t + d)$. Assuming the event from Lemma 2 occurs, note that the coordinate $J_t \in [d]$ with $z_1(J_t) = \ldots = z_{t-1}(J_t) = 1$ exists. Now, observe that any coordinate $j \in [d]$ is bound to satisfy

$$w_t(j) + w_t(j+d) \leq w_t(J_t) + w_t(J_t + d).$$

To see this, note that $w_t(J_t) = 0$, because $J_t \neq i_s$ for all $s < t$ (formally this follows by induction). In addition, by Eq. (20);

$$\forall s < t,\ z_s(J_t) = 1 \implies \forall s < t,\ -\eta g(w; z_s)(J_t + d) = \epsilon\eta \implies w_t(J_t + d) = (t-1)\epsilon\eta.$$

On the other hand, for any $j' \in [d]$ we have $w_t(j') \leq 0$, and $w_t(j' + d) \leq (t-1)\epsilon\eta$. Concluding, it follows the gradient oracle will pick $i_t = I(w_t) = J_t$, therefore $w'_{t+1}(i_t) = -\eta$ for $w'_{t+1} := w_t - \eta g_t$. To see that $w_{t+1} = \Pi_W(w'_{t+1}) = w'_{t+1}$, note that by assumption $\eta \leq 1/\sqrt{n}$, hence

$$\left\|w'_{t+1}\right\|^2 = \sum_{i=1}^{2d} w'_{t+1}(i)^2 \leq \sum_{t=1}^{t} w'_{t+1}(i_t)^2 + d(n\epsilon\eta)^2 = \eta^2(t + n/d) \leq (t+1)/n,$$

and w'_{t+1} remains inside W for all $t < n$. Finally, since the desired event occurs with probability $1/2$ by Lemma 2, we are done. □

B.1 Lower bound for large step sizes

When the step size is large, the projections actually alleviate the problematic nature of our construction, to the point where they can be exploited to obtain any convergence rate with the full iterate averaging. Notably though, concatenating our construction with a standard lower bound (e.g., Lemma 14) the best convergence rate possible is $n^{-1/4}$ with $\eta = n^{-1/4}$ which is the same as what would be achieved by the somewhat more reasonable choice of $\eta = n^{-3/4}$ that does not rely on the projections.

Theorem 9 (Large step size case of Theorem 1)**.** *Let $n \in \mathbb{N}$, $n \geq 4$, $d \geq 2^{4n \log n}$, and $W = \mathcal{B}_0^{2d}(1)$. Then there exists a distribution over instance set Z and a 4-Lipschitz convex loss function $f : W \times Z \to \mathbb{R}$ such that*

(i) the optimization error is large; $\mathbb{E}_{S \sim \mathcal{Z}^n}\left[\widehat{F}(\widehat{w}_S) - \widehat{F}(w_S^\star)\right] = \Omega\left(\frac{1}{\eta\sqrt{n}}\right)$,

(ii) the generalization gap is large; $\mathbb{E}_{S \sim \mathcal{Z}^n}\left[\widehat{F}(\widehat{w}_S) - F(\widehat{w}_S)\right] = \Omega\left(\frac{1}{\eta\sqrt{n}}\right)$,

where $\widehat{w}$ is any suffix average of SGD with step size $\eta > 1/\sqrt{n}$.

Proof. The analysis parts ways from the small step size case after Lemma 2. Instead of Lemma 3, we make the claim below.

Lemma 4. For all $\tau \in [n]$, it holds that $t < \tau \implies w_\tau(i_t) \leq -\eta(1+\eta^2)^{t-\tau}$, where $i_t := I(w_t)$ and $\tau \leq n+1$. In addition, for any suffix average $\widehat{w}$, it holds that

$$\sum_{t=n/4}^{n} |\widehat{w}(i_t)| \geq \frac{1}{20\eta}.$$

The important consequence of the above lemma is that whichever suffix average we take, we will end up with an $\Omega(1/\eta)$ mass in the total bad coordinate summation. We now show this translates to an empirical risk lower bound as claimed. Ignoring the negligible contribution of ϵ, by Lemma 15 we have;

$$\widehat{F}(w) = \frac{1}{n}\sum_{s=1}^{n} f(w; z_s) \geq \frac{1}{n}\sum_{s=1}^{n}\sqrt{\sum_{t=s+1}^{n} w(i_t)^2} \geq \frac{1}{5\sqrt{n}}\sum_{t=n/4}^{n}|w(i_t)| \geq \frac{1}{100\eta\sqrt{n}},$$

where the last inequality follows from Lemma 4. This completes the proof. □

Proof of Lemma 4. For $t \in [n]$, denote $w'_t := w_t - \eta g(w_t, z_t)$ so that now $w_{t+1} \leftarrow \Pi_W\left(w'_t\right)$. Informally, we have $\left\|w'_t\right\|^2 \leq 1 + \eta^2$ for all t, when we ignore the negligible ϵ component. Formally, let

$$\zeta_t(i) = \begin{cases} 0 & i \leq d, \\ -\epsilon z_t(i) & i > d, \end{cases}$$

so that $g(w_t; z_t) = \zeta_t + e_{i_t} + e_{(i_t+d)}$, and observe

$$\begin{aligned} \left\|w'_{t+1}\right\|^2 &= \left\|w_t - \eta\zeta_t - \eta e_{i_t} - \eta e_{(i_t+d)}\right\|^2 \\ &\leq 1 + dn\eta^2\epsilon^2 + 2\eta^2 \\ &\leq 1 + (n/d)\eta^2 + 2\eta^2 \\ &\leq 1 + 3\eta^2. \end{aligned}$$

Now, set $\gamma := 3\eta^2$, let $t < \tau \in [n]$, and observe;

$$w_\tau(i_t) = \Pi_W\left(w'_\tau\right)(i_t) = \frac{w'_\tau(i_t)}{\|w'_\tau\|} \leq w'_\tau(i_t)(1+\gamma)^{-1}$$

where the inequality follows from the norm bound and since $w'_l(i_t) \leq 0$ for all $l \in [d]$. In addition, for $\tau - 1 > t$, we have $w'_\tau(i_t) = w_{\tau-1}(i_t)$, hence

$$w'_\tau(i_t)(1+\gamma)^{-1} = w_{\tau-1}(i_t)(1+\gamma)^{-1} \leq \cdots \leq w'_{t+1}(i_t)(1+\gamma)^{t-\tau} = -\eta(1+\gamma)^{t-\tau},$$

therefore,

$$t < \tau \implies w_\tau(i_t) \leq -\eta(1+\gamma)^{t-\tau}, \tag{21}$$

which proves the first part. For the second part, we begin by computing the values in each individual coordinate.

The individual coordiantes $\widehat{w}(i_t)$. let $\widehat{w}$ be the average of the last k iterates $w_{n-k+2}, \ldots, w_{n+1}$, and set $\tau_0 := n-k$. Fix $t \in [n]$, set $l := \max(\tau_0+1, t+1)$, and observe

$$|\widehat{w}(i_t)| \geq \frac{1}{k}\sum_{\tau=l}^{n+1} |w_\tau(i_t)| \geq \frac{\eta}{k}\sum_{\tau=l}^{n+1}(1+\eta^2)^{t-\tau},$$

where the first inequality follows since all values are negative and the second from Eq. (21). We have

$$\begin{aligned}
\sum_{\tau=l}^{n+1}(1+\gamma)^{t-\tau} &= \sum_{k=l-t}^{n+1-t}(1+\gamma)^{-k} \\
&= (1+\gamma)^{-(l-t)}\left(1-(1+\gamma)^{-(n+2-l)}\right)\frac{1}{1-(1+\gamma)^{-1}} \\
&= (1+\gamma)^{t-l}\left(1-(1+\gamma)^{l-n-2}\right)\frac{(1+\gamma)}{(1+\gamma)-1} \\
&= (1+\gamma)^{t-l+1}\left(1-(1+\gamma)^{l-n-2}\right)\frac{1}{\gamma} \\
&=: (*).
\end{aligned}$$

Now,

$$\begin{aligned}
t \leq \tau_0 &\implies (*) = (1+\gamma)^{t+k-n-1}\left(1-(1+\gamma)^{-k}\right)\frac{1}{\gamma} \\
&\implies |\widehat{w}(i_t)| \geq \frac{\eta}{\gamma k}(1+\gamma)^{t+k-n-1}\left(1-(1+\gamma)^{-k}\right) \qquad (22) \\
\text{and } t > \tau_0 &\implies (*) = \left(1-(1+\gamma)^{t-n-1}\right)\frac{1}{\gamma} \\
&\implies |\widehat{w}(i_t)| \geq \frac{\eta}{\gamma k}\left(1-(1+\gamma)^{t-n-1}\right). \qquad (23)
\end{aligned}$$

Before moving on to bound the sum of values in the coordinates, we record the following basic facts which will be used repeatedly. By Bernoulli's inequality and our assumption that $\eta > 1/\sqrt{n}$, we have $(1+\gamma)^m = (1+3\eta^2)^m \geq 1+3\eta^2 m \geq 1+3m/n$. Hence,

$$\begin{aligned}
(1+\gamma)^{-m} &= \frac{1}{(1+\gamma)^m} \leq \frac{1}{1+3m/n} \\
&\implies 1-(1+\gamma)^{-m} \geq \frac{3m/n}{1+3m/n},
\end{aligned}$$

and then,

$$m \geq n/4 \implies 1-(1+\gamma)^{-m} \geq (3/4)/4 \geq 1/6, \qquad (24)$$

$$\text{and} \quad \sum_{j=0}^{m}(1+\gamma)^{-j} \geq \frac{1-(1+\gamma)^{-m}}{1-(1+\gamma)^{-1}} \geq \frac{1/6}{1-\frac{1}{1+\gamma}} = \frac{1}{6\gamma}. \qquad (25)$$

Bounding the sum of coordinate values. We first consider the case that $\tau_0 < n/2$;

$$\begin{aligned}
\sum_{t=n/4}^{n}|\widehat{w}(i_t)| &\geq \frac{\eta}{\gamma k}\sum_{t=n/2}^{3n/4}\left(1-(1+\eta^2)^{t-n-1}\right) && \text{(by Eq. (23))} \\
&\geq \frac{\eta}{\gamma k}\sum_{t=n/2}^{3n/4}\frac{1}{6} && \text{(by Eq. (24))} \\
&= \frac{\eta}{6\gamma k}(3n/4-n/2) \\
&= \frac{n\eta}{24\gamma k} \\
&\geq \frac{n}{100\eta k} \geq \frac{1}{100\eta},
\end{aligned}$$

which proves the desired result (recall that $\gamma = 3\eta^2$). Assume now $\tau_0 \geq n/2$, and observe;

$$\begin{aligned}
\sum_{t=\tau_0+1}^{n} |\widehat{w}(i_t)| &\geq \frac{\eta}{\gamma k} \sum_{t=\tau_0+1}^{n} \left(1 - (1+\gamma)^{t-n-1}\right) && \text{(by Eq. (23))} \\
&= \frac{\eta}{\gamma k} \sum_{j=1}^{k-1} \left(1 - (1+\gamma)^{-j}\right) \\
&= \frac{\eta}{\gamma k} \left(k - 1 - \sum_{j=1}^{k-1} (1+\gamma)^{-j}\right) \\
&= \frac{\eta}{\gamma k} \left(k - 1 - (1+\gamma)^{-1} \frac{1-(1+\gamma)^{1-k}}{1-(1+\gamma)^{-1}}\right) \\
&= \frac{\eta}{\gamma k} \left(k - 1 - \frac{1-(1+\gamma)^{1-k}}{\gamma}\right) \\
&= \frac{\eta}{\gamma^2 k} \left(\gamma(k-1) - (1-(1+\gamma)^{1-k})\right) \\
&= \frac{\eta}{\gamma^2 k} \left(\gamma(k-1) + (1+\gamma)^{1-k}\right). && (26)
\end{aligned}$$

In addition,

$$\begin{aligned}
\sum_{t=n/4}^{\tau_0} |\widehat{w}(i_t)| &\geq \frac{\eta}{\gamma k} \left(1 - (1+\gamma)^{-k}\right) \sum_{t=n/4}^{\tau_0} (1+\gamma)^{t+k-n-1} && \text{(by Eq. (22))} \\
&= \frac{\eta}{\gamma k} \left(1 - (1+\gamma)^{-k}\right) \sum_{t=n/4}^{\tau_0} (1+\gamma)^{t-\tau_0} && \text{(by Eq. (22))} \\
&= \frac{\eta}{\gamma k} \left(1 - (1+\gamma)^{-k}\right) \sum_{j=0}^{\tau_0-n/4} (1+\gamma)^{-j} \\
&\geq \frac{\eta}{\gamma k} \left(1 - (1+\gamma)^{-k}\right) \sum_{j=0}^{n/4} (1+\gamma)^{-j} \\
&\geq \frac{\eta}{6\gamma^2 k} \left(1 - (1+\gamma)^{-k}\right), && (27)
\end{aligned}$$

where in the last inequality we have applied Eq. (25). Now, combining Eq. (26) and Eq. (27), we obtain

$$\begin{aligned}
\sum_{t=n/4}^{n} |\widehat{w}(i_t)| &\geq \frac{\eta}{\gamma^2 k} \left(\gamma k - \gamma - 1 + (1+\gamma)^{1-k}\right) + \frac{\eta}{6\gamma^2 k} \left(1 - (1+\gamma)^{-k}\right) \\
&\geq \frac{\eta}{6\gamma^2 k} \left(\gamma k - \gamma - 1 + (1+\gamma)^{1-k} + 1 - (1+\gamma)^{-k}\right) \\
&= \frac{\eta}{6\gamma^2 k} \left(\gamma k - \gamma + (1+\gamma)^{1-k} - (1+\gamma)^{-k}\right) \\
&\geq \frac{\eta}{6\gamma k} (k-1) \\
&\geq \frac{\eta}{12\gamma} \\
&= \frac{1}{36\eta},
\end{aligned}$$

which proves the desired result also in the second case, and completes the proof. □

B.2 SCO with non convex components

Proof of Theorem 2. We define the distribution $\mathfrak{Z} = \mathfrak{Z}(\delta)$ over the set of datapoints Z by

$$\forall i \le d; \quad z(i) = \begin{cases} -1 & \text{w.p. } \delta \\ 1 & \text{w.p. } \delta \\ 0 & \text{w.p. } (1-2\delta), \end{cases}$$
$$\forall i > d; \quad z(i) = \mathbb{1}\left\{z(i-d) = -1\right\}.$$

We consider the same loss function of Theorem 1, but leave the norm-like component without the square-root;

$$f(w;z) := \phi(w;z) + \nu_z(w), \quad \nu_z(w) := \sum_{i=1}^{d} z(i) w(i)^2,$$

where ϕ is defined as in Eq. (13). We also define the gradient oracle for ϕ as we have done in the convex case Eq. (15), Eq. (16), and Eq. (17), repeated here for convenience;

$$g_\phi(w;z)(i) := \begin{cases} \mathbb{1}\left\{i = I(w)\right\} & i \le d \\ -\epsilon z(i) + \mathbb{1}\left\{i = I(w) + d\right\} & i \ge d \end{cases},$$
$$I(w) := \min\left\{i \in [d] \mid i \in \arg\max_{1 \le j \le d}\{w(j) + w(j+d)\}\right\},$$
$$i_t := I(w_t).$$

Here, unlike the construction in Theorem 1 we need ϵ to depend on η, and set $\epsilon := \eta/d$. The next lemma establishes the SGD iterates end up "overfitting" the empirical objective, and follows from a proof that is essentially identical to Lemma 2 and Lemma 3. The only difference is that here the training examples have -1 rather than 1 in the critical coordinates.

Lemma 5. For $\delta = 1/4n^2$, we have with probability $\ge 1/2$ that for all $\tau \in [n]$ and $t < \tau$;

- $w_\tau(i_t) = -\eta$,
- $1 \le s < t \implies z_s(i_t) = -1$, *and*
- $t \le s \le n \implies z_s(i_t) \le 0$.

Thus, let $\tau \le n+1$, and denote $\overline{w}_\tau := \frac{1}{n-\tau+2}\sum_{t=\tau}^{n+1} w_t$. By a derivation identical to that of the convex case, we obtain for all $1 \le t \le n/2$, $\overline{w}_\tau(i_t) \le -\frac{\eta}{2}$. Hence, for $s \le n/4$,

$$\nu_{z_s}(\overline{w}_\tau) \le -\sum_{t=n/4+1}^{n/2} \overline{w}_\tau(i_t)^2 = -\frac{n\eta^2}{16},$$

therefore

$$\begin{aligned} \widehat{F}(\overline{w}_\tau) &= \frac{1}{n}\sum_{t=1}^{n} \phi(\overline{w}_\tau; z_t) + \frac{1}{n}\sum_{t=1}^{n} \nu_{z_t}(\overline{w}_\tau) \\ &\le 2\epsilon\eta n - \frac{1}{n}\frac{n}{2}\frac{n\eta^2}{16} \\ &\le -\frac{n\eta^2}{64}. \end{aligned}$$

To conclude the proof, we note that

$$F(\overline{w}_\tau) \ge -\epsilon \sum_{i=d+1}^{2d} \delta(\eta\epsilon n) = -\epsilon^2 d\delta n\eta \ge -\eta^2/n,$$

and the result follows. □

B.3 SCO with strongly convex components

Proof of Theorem 3. We first make the argument for an unbounded domain, so that no projections take place. Let

$$f(w;z) := \phi(w;z) + \nu_z(w) + \frac{\lambda}{2}\|w\|^2,$$

where ϕ and ν are defined by

$$\phi(w;z) := -\epsilon \sum_{i=d+1}^{2d} z(i)w(i) + \max_{1\le i\le d}\{w(i) + \epsilon w(i+d)\}, \tag{28}$$

$$\nu_z(w) := \sqrt{\sum_{i=1}^{d} z(i)w(i)^2}, \tag{29}$$

and $\epsilon := 1/d$. These are essentially the same definitions as in our main construction Eq. (13) and Eq. (14), but with an added ϵ factor inside the max component of ϕ. This only makes our formal argument simpler, but otherwise does not make any significant difference. For the gradient oracle, we define

$$g_\phi(w;z)(i) := \begin{cases} \mathbb{1}\{i = I(w)\} & i \le d \\ -\epsilon z(i) + \epsilon\mathbb{1}\{i = I(w) + d\} & i \ge d \end{cases}, \tag{30}$$

$$\text{where } I(w) := \min\left\{i \in [d] \mid i \in \arg\max_{1\le j\le d}\{w(j) + \epsilon w(j+d)\}\right\}, \tag{31}$$

and again denote the index picked by g on round $t \in [n]$ by

$$i_t := I(w_t). \tag{32}$$

We then set

$$g(w;z) := g_\phi(w;z) + \nabla\nu_z(w) + \lambda w.$$

Clearly, for all $w, z \in \mathbb{R}^{2d}$, $g(w;z) \in \partial_w f(w;z)$. Following a direct computation, we get that

$$g(w;z)(i) = \begin{cases} \mathbb{1}\{i = I(w)\} + \frac{z(i)w(i)}{\nu_z(w)} + \lambda w(i) & i \le d, \\ \epsilon\mathbb{1}\{i = I(w) + d\} - \epsilon z(i) + \lambda w(i) & i > d, \end{cases}$$

where I is defined in Eq. (16). Hence, the stochastic gradient steps $w_{t+1} \leftarrow w_t - \eta_t g(w_t, z_t)$ are given by

$$w_{t+1}(i) = \begin{cases} \left(1 - \eta_t(z_t(i)\nu_{z_t}(w)^{-1} + \lambda)\right) w_t(i) - \eta_t\mathbb{1}\{i = I(w)\} & i \le d \\ (1 - \eta_t\lambda)\, w_t(i) + \epsilon\eta_t z_t(i) - \eta_t\epsilon\mathbb{1}\{i = I(w) + d\} & i > d. \end{cases} \tag{33}$$

The next lemma makes a similar assertion as Lemma 3 and follows from similar arguments.

Lemma 6. With probability $\ge 1/2$, for all $\tau \in [n]$ we have

$$w_{\tau+1}(i) = \begin{cases} -\eta_t \prod_{s=t+1}^{\tau}(1 - \eta_s\lambda) & i = i_t, t \in [\tau], \\ 0 & i \in [d] \setminus \{i_1, \dots i_\tau\}, \end{cases}$$

where $z_1(i_t) = \dots = z_{t-1}(i_t) = 1$, and $z_t(i_t) = \dots = z_n(i_t) = 0$.

Next, a simple derivation shows the empirical risk is large for any suffix average. Thus, let $\tau \le n+1$, and denote $\overline{w}_\tau := \frac{1}{n-\tau+2}\sum_{t=\tau}^{n+1} w_t$. By Lemma 6, assuming the event from the lemma occurs we have

$$t \le n/2 \implies |\overline{w}_\tau(i_t)| \ge \frac{1}{2}|w_{n+1}(i_t)| = \frac{\eta_t}{2}\prod_{s=t+1}^{n}(1 - \eta_s\lambda).$$

Therefore, noting that $\phi(\overline{w}_\tau; z_s) \ge -\epsilon^2 dn/\lambda \ge -\epsilon n/\lambda \ge -1/(\lambda n)$, by Lemma 15 we obtain;

$$\begin{aligned} \widehat{F}(\overline{w}_{\tau:n}) &\ge \frac{\lambda}{2}\|\overline{w}_{\tau:n}\|^2 - \frac{1}{\lambda n} + \frac{1}{10\sqrt{n}}\sum_{t=n/4}^{n/2}\eta_t\prod_{s=t+1}^{n}(1 - \eta_s\lambda) \\ &\ge \frac{\lambda}{2}\|\overline{w}_{\tau:n}\|^2 - \frac{1}{\lambda n} + \frac{1}{40\lambda\sqrt{n}}. \qquad \text{(for } \eta_t = 1/\lambda t) \end{aligned}$$

Noting that $\widehat{F}(0) = 0$, we obtain the claim on the optimization error. For the generalization gap, first note that for any $t \leq n+1$,

$$\begin{aligned}
\mathbb{E}_z \phi(w_t; z) &\leq -\epsilon \sum_{i=d+1}^{2d} \delta w_t(i) + \epsilon n/\lambda \\
&\leq \epsilon \delta n/\lambda + \epsilon n/\lambda \\
&\leq 2\epsilon n/\lambda,
\end{aligned}$$

and observe;

$$\begin{aligned}
F(\overline{w}_\tau) &\leq \frac{2\epsilon n}{\lambda} + \frac{\lambda}{2} \|\overline{w}_\tau\|^2 + \mathbb{E}_{z\sim\mathfrak{Z}} \left[\sqrt{\sum_{t=1}^{n} z(i_t) \overline{w}_\tau(i_t)^2} \right] \\
&\leq \frac{2\epsilon n}{\lambda} + \frac{\lambda}{2} \|\overline{w}_\tau\|^2 + \sqrt{\delta \sum_{t=1}^{n} \overline{w}_\tau(i_t)^2} && \text{(by Jensen's inequality)} \\
&\leq \frac{2\epsilon n}{\lambda} + \frac{\lambda}{2} \|\overline{w}_\tau\|^2 + \frac{1}{2n} \sum_{t=1}^{n} |\overline{w}_\tau(i_t)| \\
&\leq \frac{2\epsilon n}{\lambda} + \frac{\lambda}{2} \|\overline{w}_\tau\|^2 + \frac{\log n}{2\lambda n} && \text{(for } \eta_t = 1/\lambda t) \\
&\leq \frac{\lambda}{2} \|\overline{w}_\tau\|^2 + \frac{\log n}{\lambda n}. && \text{(since } 2\epsilon n/\lambda \leq \log n/(2\lambda n))
\end{aligned}$$

Combining the inequalities in the last two displays completes the proof.

Bounded domain case with $\lambda \geq 1/\sqrt{n}$. In this case projections happen, but owed to our assumption on λ we will see their effect is negligible. Denote $w'_s := w_s - \eta g(w_s, z_s)$ so that now $w_{s+1} \leftarrow \Pi_W\left(w'_s\right)$. Observe that by Eq. (33) under the event of Lemma 6, we have

$$w'_{\tau+1}(i) = \begin{cases} 0 & i \in [d] \setminus \{i_1, \ldots, i_\tau\}, \\ (1-\eta_\tau\lambda) w_\tau(i) + \eta_\tau \mathbb{1}\{i = i_\tau\} & i \in \{i_1, \ldots, i_\tau\}, \\ (1-\eta_\tau\lambda) w_\tau(i) + \epsilon\eta_\tau z_t(i)(-1 + \mathbb{1}\{i = i_\tau + d\}) & i \in [d+1, 2d], \end{cases}$$

thus

$$w'_{\tau+1}(i)^2 \leq \begin{cases} 0 & i \in [d] \setminus \{i_1, \ldots, i_\tau\}, \\ (1-\eta_\tau\lambda)^2 w_\tau(i)^2 & i \in \{i_1, \ldots, i_{\tau-1}\}, \\ \eta_\tau^2 & i = i_\tau, \\ (n\epsilon/\lambda)^2 & i \in [d+1, 2d]. \end{cases}$$

Note that by our assumption that $\lambda \geq 1/\sqrt{n}$ and our choice of $\epsilon = 1/d$, we have $n^2\epsilon^2/\lambda^2 \leq n/d^2$. Now, fix $\tau \geq 2n/3$, and observe;

$$\begin{aligned}
\left\|w'_{\tau+1}\right\|^2 &\leq \eta_\tau^2 + d(n/d^2) + (1-\eta_\tau\lambda)^2 \sum_{i=1}^{2d} w_\tau(i)^2 \\
&= \eta_\tau^2 + n/d + (1-\eta_\tau\lambda)^2 \|w_\tau\|^2 \\
&\leq \eta_\tau^2 + n/d + (1-\eta_\tau\lambda)^2 \\
&= \frac{1}{\lambda^2\tau^2} + \frac{n}{d} - \frac{2}{\tau} + \frac{1}{\tau^2} \\
&\leq \frac{2}{\lambda^2 n\tau} + 1 + \frac{n}{d} + \frac{9}{4n^2} - \frac{2}{\tau} \\
&\leq \frac{3}{2\tau} + 1 + \frac{10}{4n^2} - \frac{2}{\tau} \\
&\leq 1.
\end{aligned}$$

In the above, we have used that $n/d \leq 1/4n^2$, and that $1/2\tau \geq 1/(2n) \geq 10/4n^2$ for $n \geq 10$. Hence, from round $2n/3$ onwards projections do not occur anymore. To conclude the proof, we note that we can lower bound the empirical loss precisely as we did before but over rounds $4n/6$ to $5n/6$, rather than $n/4$ to $n/2$. In addition, the population loss has only improved since the per coordinate values in all iterates have only decreased in magnitude as a result of the projections. □

Proof of Lemma 6. Note that for all $t \in [n]$ and any $i \in [d+1, 2d] \setminus \{i_1 + d, \ldots, i_t + d\}$, by Eq. (33) we have

$$\begin{aligned} w_{t+1}(i) &= (1 - \eta_t \lambda)\, w_t(i) + \epsilon \eta_t z_t(i) \\ &= \sum_{s=1}^{t} \epsilon \eta_s z_s(i) \prod_{l=s+1}^{t} (1 - \eta_l \lambda)\,. \end{aligned} \tag{34}$$

Hence, the number of times $z_s(i) = 1$ for $s \leq t$ determines the maximality of $w_{t+1}(i)$. In other words, the extra component in the gradient update effects all coordinates equally, and the situation here is no different than the convex case. Thus, by Lemma 2 and the same arguments as given in Lemma 3, we have that with probability $\geq 1/2$, for all $t \in [n]$, $i_t = I(w_t) = J_t$. Therefore, by Eq. (33);

$$w_{t+1}(i) = \begin{cases} -\eta_t & i = i_t; \\ -\eta_s \prod_{l=s+1}^{t} (1 - \eta_l \lambda) & i = i_s,\ s < t; \\ 0 & i \in [d] \setminus \{i_1, \ldots, i_t\}\,. \end{cases}$$

Note that the $z_t(i) v_{z_t}(w)^{-1}$ component in Eq. (33) does not contribute since $z_t(i_s) = 0$ for all $t \geq s$, given our event. □

C Proofs for Section 4

In what follows we provide the standard analysis of SGD with the iterate averaging scheme specified in Theorem 4. The theorem stated and proved below provides the rate of convergence on the target objective function from which gradients are sampled (as similar analyses normally do); note that we use it in the context where the target objective is the *empirical* loss given by the training set. This should be contrasted with the goal of Theorem 4, which is to establish the convergence rate on the *population* objective. Our only motivation for proving the below theorem is to argue the generalization gap upper bound established in Corollary 2.

Theorem 10. *Let $W \subset \mathbb{R}^d$ with diameter D, and $f_1, \ldots, f_n$ be a sequence of convex, G-Lipschitz losses sampled i.i.d. from some distribution $\mathcal{F}$. Further, let $w^\star := \min_{w \in W} \mathbb{E}_{f \sim \mathcal{F}} f(w)$ denote the minimizer of the expected function. Then, the weighted average $\overline{w} := \frac{2}{n+1} \sum_{t=1}^{n} \frac{n-t+1}{n} w_t$ of the iterates produced by SGD with step size $\eta = \frac{D}{G\sqrt{n}}$ obtains the following upper bound:*

$$\mathbb{E}_{f_1, \ldots, f_n, f \sim \mathcal{F}} \left[f(\overline{w}) - f(w^\star) \right] \leq \frac{4GD}{\sqrt{n}}.$$

Proof. Observe,

$$\begin{aligned} \mathbb{E}\left[f(\overline{w}) - f(w^\star) \right] &\leq \frac{2}{n+1} \sum_{t=1}^{n} \frac{n-t+1}{n} \mathbb{E}\left[f(w_t) - f(w^\star) \right] \\ &= \frac{2}{n+1} \mathbb{E}\left[\sum_{t=1}^{n} \frac{n-t+1}{n} \left(f_t(w_t) - f_t(w^\star) \right) \right]. \end{aligned} \tag{35}$$

By the standard SGD analysis,

$$f_t(w_t) - f_t(w^\star) \leq \nabla f_t(w_t)^\top (w_t - w^\star) \leq \frac{1}{2\eta} \left(D_t^2 - D_{t+1}^2 \right) + \frac{\eta}{2} G^2,$$

where $D_t := \|w_t - w^\star\|$. Now,

$$\begin{aligned}
\sum_{t=1}^{n} \frac{n-t+1}{n}\left(f_t(w_t) - f_t(w^\star)\right) &\le \frac{1}{2\eta}\sum_{t=1}^{n}\frac{n-t+1}{n}\left(D_t^2 - D_{t+1}^2\right) + \frac{\eta G^2 n}{2} \\
&= \frac{D_1^2}{2\eta} + \frac{1}{2\eta}\sum_{t=2}^{n} D_t^2\left(\frac{n-t+1}{n} - \frac{n-t}{n}\right) + \frac{\eta G^2 n}{2} \\
&= \frac{D_1^2}{2\eta} + \frac{1}{2\eta n}\sum_{t=2}^{n} D_t^2 + \frac{\eta G^2 n}{2} \\
&\le \frac{D^2}{\eta} + \frac{\eta G^2 n}{2} \\
&\le 2GD\sqrt{n}.
\end{aligned}$$

Plugging the above inequality into Eq. (35), we obtain

$$\mathbb{E}\left[f(\overline{w}) - f(w^\star)\right] \le \frac{4GD}{\sqrt{n}},$$

which completes the proof. □

D Proofs for Section 5

D.1 Lower bound for multi-epoch SGD

Proof of Theorem 5. First, note that without modifications, the strategy of Theorem 1 breaks after the first epoch; it will just keep pointing the gradient on coordinates with an all ones sequence. We use the idea we can fully "record" into the iterate w_t^k the precise samples we have stepped through so far, and define a gradient oracle that will cause the iterate to advance on fresh bad coordinates in every new epoch. We will work with the datapoints set $Z = \{0,1\}^d$ and define the distribution $\mathcal{Z} = \mathcal{Z}(\delta)$ over Z by letting $z(i) \sim \mathrm{Ber}(\delta)$ for all $i \in [d]$. We consider two separate portions of a vector $w \in \mathbb{R}^{d'}$, which we denote by $w[\cdot;0] \in \mathbb{R}^d$ and $w[\cdot,\cdot;1] \in \mathbb{R}^{d\times nK}$. The first portion with d entries is where the bad gradient steps will be made and where we will eventually suffer the loss from. The second consists of dnK entries and is used to encode the samples observed during the optimization process. Our loss function is defined as follows;

$$\begin{aligned}
v_z(w) &:= \sqrt{\sum_{i=1}^{d} z(i) w[i;0]^2}, \\
\phi(w;z) &:= \epsilon \sum_{i=1}^{d}(1+z(i)) \max_{j\in[nK]}\{w[i,j;1]\} + \max_{i\in[d]}\{w[i;0])\}, \\
f(w;z) &:= \phi(w;z) + v_z(w).
\end{aligned} \tag{36}$$

Again, we choose $\epsilon > 0$ sufficiently small so that the loss induced by it is negligible, and so that f is 4-Lipschitz. The gradient oracle we use is specified by that of the ϕ function;

$$\begin{aligned}
g_\phi(w;z)[i,j;1] &= \epsilon(1+z(i))\mathbb{1}\{j = I_1(w[i,\cdot;1])\} \\
g_\phi(w;z)[\cdot;0] &= e_{I_0(w)} \\
\text{where } I_1(x) &:= \min\left\{ j \in [nK] \;\middle|\; j \in \operatorname*{arg\,max}_{l\in[nK]}\{x(l)\}\right\} \quad \text{for } x \in \mathbb{R}^{nK}, \\
I_0(w) &:= \min\left\{ i \in [d] \;\middle|\; i \in \operatorname*{arg\,max}_{l\in[d]}\{w[l;0]\},\ \text{and } i \in \operatorname*{arg\,min}_{l\in[d]}\left\{V(w)(l)\right\}\right\}, \\
\text{and } V(w) &:= \sum_{j=I_0(w)}^{nK} w[\cdot,j;1],
\end{aligned}$$

where $t_0(w)$ denotes the first global iteration index of the current epoch. This index is easy to infer from $w[\cdot,\cdot;1]$ since the t'th SGD iteration in epoch k results in values strictly smaller than 0 in all entries of $w[\cdot,\tau;1]$, where $\tau = n(k-1)+t$ (and 0 remains from initialization for $\tau' > \tau$). In words, we design our "adversarial" gradient oracle so that it will choose the coordinate for $\max_{i\in[d]}\{w[i;0]\}$, by "looking" in $w[\cdot,\cdot;1]$ and choosing the coordinate $i_t^k := I_0(w_t^k) \in [d]$ such that the number if times $z_s^k(i_t^k) = 1$ for $s < t$ is largest. An illustration is provided in Fig. 1.

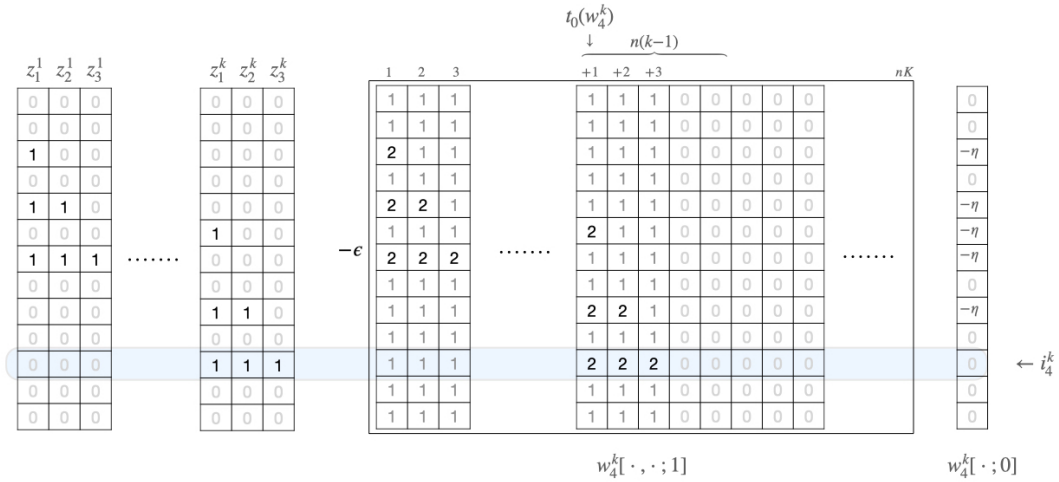

Figure 1: Illustration of gradient oracle mechanism

In similar spirit to the the basic construction from Theorem 1, we will ensure that with high probability, the coordinates selected by our gradient oracle are such that $z_s^k(i_t^k) = 1$ for all $s < t$, and $z_s^k(i_t^k) = 0$ for all $s \geq t$. To that end, we first assert the existence of a set of datapoints $Z \subset \{0,1\}^d$ where a desired property described next holds with sufficiently high probability. Consider some arbitrary ordered set $S = \{z_1,\ldots,z_n\} \subset \{0,1\}^d$. For $t \in [n]$, denote

$$E^{(1,t)} := \left\{i \in [d] \mid z_s(i) = 1\ \forall s < t\right\},$$
$$E^{(1,t;K)} := \left\{J_t^1,\ldots,J_t^K \mid J_t^k \text{ is the k'th smallest} \in E^{(1,t)}\right\},$$
$$E^{(t,0)} := \left\{i \in [d] \mid z_s(i) = 0\ \forall s \geq t\right\}.$$

So $E^{(1,t;K)}$ is just the first K elements of $E^{(1,t)}$, where we enumerate the coordinates by the superscript in increasing order. We say the event $\mathcal{E}$ holds for the set S, or equivalently that $S \in \mathcal{E}$ if for all $t \in [n]$, we have $|E^{(1,t;K)}| = K$, and $E^{(1,t;K)} \subseteq E^{(t,0)}$. In words, $S \in \mathcal{E}$ if for every $t \in [n]$, the first K coordinates $\left\{J_t^1,\ldots,J_t^K\right\}$ that have a prefix of $(t-1)$ ones; $s < t \implies z_s(J_t^k) = 1$, also satisfy that they have a suffix of zeros; $s \geq t \implies z_s(J_t^k) = 0$.

Lemma 7. There exists a set of datapoints $Z = \{\zeta_1,\ldots,\zeta_n\} \subset \{0,1\}^d$, such that

$$\Pr_{\pi_1,\ldots,\pi_K\sim\Pi([n])}\left(\forall k \in [K],\ \left\{z_1^k,\ldots,z_n^k\right\} \in \mathcal{E}\right) \geq 1/2,$$

where π_k is sampled by either single-shuffle or multi-shuffle, and $z_t^k := \zeta_{\pi_k(t)}$.

With Lemma 7 in place, we can be sure bad coordinate sets will turn up in every epoch. Let $k \in [K]$ and $t \in [n]$, and assume the event from the lemma occurs. By the definition of our gradient oracle, it is bound to select one of the first k coordinates that have a prefix of all ones, which we are assured by the lemma will also have a suffix of zeros. Formally, we argue by induction on k,t. The base case follows from the definition of g_ϕ and our assumption that the event occurs. For the inductive step, assume the selected coordinates $i_{t'}^{k'}$ of all prior rounds satisfy the inductive hypothesis. Then at most $k-1$ of the first coordinates in $E^{(1,t,K)}$ could have been selected previously, since the inductive hypothesis implies every selected coordinate $i_{t'}^{k'}$ has exactly $t'-1$ ones. Hence, the gradient oracle

will choose a coordinate from $E^{(1,t,K)}$, the elements of which are coordinates that also enjoy a suffix of zeros, as assured by the event from the lemma. Summarizing, we have that for every $k \in [K]$, $t \leq n+1$, it holds that:

$$\begin{aligned} t < t' &\implies w^k_{t'}(i^k_t) = -\eta \\ \text{and } s < t &\implies z^k_s(i^k_t) = 1. \end{aligned}$$

To complete the proof, we will now prove a lower bound of $\Omega(\eta\sqrt{n/J})$ for the average iterate of the last J epochs. The other terms in the bounds of the theorem statement follow from concatenating our problem instance dimension-wise with standard constructions — see Lemma 14. Proceeding, we slightly overload notation and denote $w(i)$ for $w[i,0]$. Let $\widehat{w}$ be the average of the iterates in the last $J \in [K]$ epochs;

$$\widehat{w} := \frac{1}{nJ} \sum_{k=K-J+1}^{K} \sum_{t=1}^{n+1} w^k_t.$$

For all $n/4 \leq t \leq 3n/4$ and $K - J + 1 \leq k \leq K$, we have

$$\left|\widehat{w}(i^k_t)\right| \geq \frac{1}{nJ} \sum_{t'=3n/4}^{n} \left|w^k_{t'}(i^k_t)\right| \geq \frac{\eta}{nJ}(n/4) = \frac{\eta}{4J}, \tag{37}$$

since $w^k_{t'}(i^k_t) = -\eta$ for all $t' \geq t$.

Single-shuffle case. Ignoring the negligible ϵ terms, we now have

$$\begin{aligned} 1 \leq s \leq n/4 \implies f(\widehat{w}; z^1_s) &\geq \sqrt{\sum_{k=1}^{K} \sum_{t=s}^{n/2} z^1_s(i^k_t)\widehat{w}(i^k_t)^2} \\ &\geq \sqrt{\sum_{k=K-J+1}^{K} \sum_{t=n/4+1}^{n/2} z^1_s(i^k_t)\widehat{w}(i^k_t)^2} = \sqrt{(nJ/4)\frac{\eta^2}{16J^2}} = \sqrt{\frac{n}{64J}}\eta, \end{aligned}$$

since $z^k_s(i^k_t) = z^1_s(i^k_t) = 1 \ \forall s < t,\ k \in [K]$. Therefore

$$F(\widehat{w}) = \frac{1}{n}\sum_{s=1}^{n} f(\widehat{w}; z^1_s) \geq \frac{n/4}{n}\sqrt{\frac{n}{64J}}\eta \geq \sqrt{\frac{n}{J}}\frac{\eta}{32},$$

concluding the proof for this case.

Multi-shuffle case. For $\zeta_i \in Z$ denote

$$\begin{aligned} I^k(\zeta_i) &:= \mathbb{1}\left\{z_t = \zeta_i, t \in [n/4, 3n/4]\right\}, \\ I(\zeta_i) &:= \sum_{k=K-J+1}^{K} I^k(\zeta_i), \\ \widetilde{Z} &:= \left\{\zeta_i \in Z \mid I(\zeta_i) \geq J/4\right\}. \end{aligned}$$

We wish to lower bound the size of $\widetilde{Z}$, to show that enough ζ_i's where incident in the $[n/4, 3n/4]$ iteration range in a sufficiently large number of epochs. (Our interest in this range stems from the desire to apply Eq. (37).) Observe;

$$\begin{aligned} &\sum_{k=K-J+1}^{K} \sum_{i=1}^{n} I^k(\zeta_i) = J\frac{n}{2} \\ &\implies \sum_{i=1}^{n} I(\zeta_i) = \frac{nJ}{2}. \end{aligned}$$

Since $I(\zeta_i) \leq J$ for all $i \in [n]$, by the pegionhole principle we get that $\left|\widetilde{Z}\right| \geq n/4$. Otherwise, we would have

$$\sum_{i=1}^{n} I(\zeta_i) \leq \left|\widetilde{Z}\right|J + \left(n - \left|\widetilde{Z}\right|\right)\frac{J}{4} < J\frac{n}{4} + \frac{J}{4}n = \frac{nJ}{2}.$$

Now,

$$
\begin{aligned}
F(\widehat{w}) \geq \frac{1}{n}\sum_{z\in\widetilde{Z}} f(\widehat{w};z) &\geq \frac{1}{n}\sum_{z\in\widetilde{Z}}\sqrt{\sum_{k=K-J+1} I^k(z)\sum_{t=3n/4}^{n}\widehat{w}(i_t^k)^2} \\
&\geq \frac{1}{n}\sum_{z\in\widetilde{Z}}\sqrt{\frac{J}{4}\sum_{t=3n/4}^{n}\frac{\eta^2}{16J^2}} \\
&= \frac{1}{n}\sum_{z\in\widetilde{Z}}\sqrt{\frac{n\eta^2}{16^2 J}} \\
&\geq \frac{\eta\sqrt{n}}{64\sqrt{J}},
\end{aligned}
$$

which concludes the multi-shuffle case and the proof as a whole. □

Proof of Lemma 7. We will make our argument for an i.i.d. sampled instance set Z, and convert it to the stated result as follows. Assume $\mathfrak{Z}$ is a distribution over $\{0,1\}^d$ for which we establish the following;

$$
\Pr_{Z\sim\mathfrak{Z}^n}(Z\in\mathcal{E}) \geq \frac{1}{2^{1/K}}. \tag{38}
$$

Clearly, applying a random permuation on $Z\sim\mathfrak{Z}^n$ does not change its distribution, therefore

$$
\begin{aligned}
\frac{1}{2^{1/K}} \leq \Pr_{Z\sim\mathfrak{Z}^n}(Z\in\mathcal{E}) &= \Pr_{Z\sim\mathfrak{Z}^n,\pi\sim\Pi([n])}(\pi(Z)\in\mathcal{E}) \\
&= \sum_{Z\in Z^n}\Pr_{\mathfrak{Z}^n}(Z)\Pr_{\pi\sim\Pi([n])}(\pi(Z)\in\mathcal{E}) \\
&\leq \max_{Z\in Z^n}\left\{\Pr_{\pi\sim\Pi([n])}(\pi(Z)\in\mathcal{E})\right\}.
\end{aligned}
$$

The above derivation implies the existence of $Z^\star\in Z^n$ with the property that

$$
\Pr_{\pi_1,\ldots,\pi_K\sim\Pi([n])}(\forall k\leq K,\ \pi_k(Z^\star)\in\mathcal{E}) \geq \left(\frac{1}{2^{1/K}}\right)^K = \frac{1}{2}, \tag{multi-shuffle}
$$

$$
\text{and } \Pr_{\pi_1\sim\Pi([n])}(\forall k\leq K,\ \pi_k(Z^\star)\in\mathcal{E}) \geq \frac{1}{2^{1/K}} \geq \frac{1}{2},\ \text{ where } \pi_1=\ldots=\pi_K. \tag{single-shuffle}
$$

Therefore, for the rest of the proof we focus on proving the distribution $\mathfrak{Z}$ as defined next satisfies the desired property Eq. (38). Let $\delta>0$ which will be chosen in hindsight, and consider $\mathfrak{Z}=\mathfrak{Z}(\delta)$ where $z(i)\sim\mathrm{Ber}(\delta)$ for each $i\in[d]$ independently. Fix $t\in[n]$, and let $\mathcal{E}_t$ denote the event that $|E^{(1,t)}|\leq K$, and $E^{(1,t;K)}\subseteq E^{(t,0)}$. We will prove $\mathcal{E}_t$ holds with sufficiently high probability, so that $\mathcal{E}=\cap_{t\in[n]}\mathcal{E}_t$ holds w.p. $\geq 1/2^{1/K}$. Proceeding, assuming we choose δ and d so that $K<d\delta^n$, by Hoeffding's inequality we have that

$$
\begin{aligned}
\Pr\left(|E^{(1,t)}|\leq K\right) &= \Pr\left(\sum_{i=1}^{d}\mathbb{1}\left\{i\in E^{(1,t)}\right\}\leq K\right) \\
&= \Pr\left(\sum_{i=1}^{d}\mathbb{1}\left\{i\in E^{(1,t)}\right\} - d\delta^{t-1}\leq K-d\delta^{t-1}\right) \\
&= \Pr\left(d\delta^{t-1}-\sum_{i=1}^{d}\mathbb{1}\left\{i\in E^{(1,t)}\right\}\geq d\delta^{t-1}-K\right) \\
&\leq e^{-(d\delta^{t-1}-K)^2/d} \leq e^{-(d\delta^n-K)^2/d}.
\end{aligned} \tag{39}
$$

In addition, for $i\in\left\{J_t^1,\ldots,J_t^K\right\}=E^{(1,t;K)}$, we have

$$
\Pr(i\in E^{(t,0)}) = \Pr(\forall s\geq t,\ z_s(i)=0) \geq (1-\delta)^n.
$$

Therefore,

$$\Pr(E^{(1,t;K)} \subseteq E^{(t,0)}) \geq (1-\delta)^{nK}.$$

From the above and Eq. (39) we obtain

$$\Pr(\text{not } \mathcal{E}_t) = \Pr\left(E^{(1,t;K)} \nsubseteq E^{(t,0)} \text{ or } |E^{(1,t)}| < K\right) \leq e^{-(d\delta^n - K)^2/d} + 1 - (1-\delta)^{nK},$$

hence,

$$\Pr\left(\cap_{t\in[n]}\mathcal{E}_t\right) \geq 1 - n(e^{-(d\delta^n-K)^2/d} + 1 - (1-\delta)^{nK}). \tag{40}$$

To finish the proof, we choose δ and d as follows. Set $\delta := 1/cn^2K$, and note that

$$(1-\delta)^{nK} = \left(1 - \frac{1}{cn^2K}\right)^{nK} \geq 1 - \frac{1}{cn} \implies 1-(1-\delta)^{nK} \leq \frac{1}{cn}.$$

In addition, note that $-(d\delta^n - K)^2/d \leq -d\delta^{2n} + 2K\delta^n$, hence

$$\begin{aligned} e^{-(d\delta^n-K)^2/d} &\leq \frac{1}{cn} \\ \Longleftarrow\ e^{-d\delta^{2n}+2K\delta^n} &\leq \frac{1}{cn} \\ \Longleftrightarrow\ \log(cn) + 2K\delta^n &\leq d\delta^{2n} \\ \Longleftrightarrow\ (\log(cn) + 2K\delta^n)(cn^2K)^{2n} &\leq d, \end{aligned}$$

which holds for any $d \geq 2^{6n\log(cnK)} = (cnK)^{6n} \geq 2\log(cn)(cn^2K)^{2n}$. Back to Eq. (40) we obtain for $c = \frac{4}{2^{1/K}-1}$;

$$\Pr\left(\cap_{t\in[n]}\mathcal{E}_t\right) \geq 1 - n\left(\frac{1}{cn} + \frac{1}{cn}\right) = 1 - \frac{2}{c} \geq \frac{1}{2^{1/K}},$$

and we are done. □

D.2 Upper bound for single-shuffle multi-epoch SGD

First, we slightly generalize the notion of uniform argument stability and prove some supporting lemmas. We extend the definition of uniform-argument-stability Eq. (4) to one that enables more than one difference in the training sets. We give the definition below in notation suitable for SGD and the lemmas that follow;

$$\epsilon_{\text{stab}}^{\text{SGD}}(\tau; J) := \max_{f_1,\ldots,f_\tau,f'_1,\ldots,f'_J} \max_{i_1,\ldots,i_J\in[\tau]} \left\|w_{\tau+1} - w'_{\tau+1}\right\|, \tag{41}$$

where $w'_{\tau+1}$ is the output of GD after swapping $f_1,\ldots,f_\tau$ in locations $i_1,\ldots,i_J$ with the other losses $f'_1,\ldots,f'_J$. Lemma 8 given next generalizes Lemma 10 for the stability notion we have introduced above. The proof provided below is based on similar lemmas given in [20].

Lemma 8. *Let $\{f(w;t)\}_{t=1}^n$ be a set of n, G-Lipschitz losses, and $F(w) = \frac{1}{n}\sum_{t=1}^n f(w;t)$. Then, for a uniformly random permutation $\pi\colon [n] \leftrightarrow [n]$, $f_t^k = f(\cdot;\pi(t))\forall k$, it holds that for single-shuffle SGD;*

$$\mathbb{E}\left[F(w_t^k) - f_t^k(w_t^k)\right] \leq G\epsilon_{\text{stab}}^{\text{SGD}}(n(k-1)+t-1; 2k),$$

where w_t^k the t'th SGD iterate of the k'th epoch.

Proof. Fix $t, i \in [n]$, and let $\pi(f_t \leftarrow f(\cdot;i))$ denote the distribution obtained from a random permutation followed by replacing f_t with $f(\cdot;i)$. In addition, denote by $\pi \mid f_t = f(\cdot;i)$ a uniformly distributed permutation conditioned on $f_t = f(\cdot;i)$. It is easily verified both distributions coincide.

Now, by the law of total expectation;

$$\begin{aligned}
\mathbb{E}_{f_1 \dots f_n}\left[f_t(w_t^k)\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_1 \dots f_n}\left[f(w_t^k; i) \mid f_t = f(\cdot; i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_1 \dots f_n \sim \pi \mid f_t = f(\cdot; i)}\left[f(w_t^k; i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_1 \dots f_n \sim \pi(f_t \leftarrow f(\cdot; i))}\left[f(w_t^k; i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_1 \dots f_n}\left[f(w_t^{k,(i)}; i)\right],
\end{aligned}$$

where $w_t^{k,(i)}$ denotes the SGD iterate obtained for the datapoint sequence after replacing f_t^j with $f(\cdot; i)$ in all epochs $j \leq k$. Note this means each epoch differs from its original version in either 0 or 2 indexes. Now

$$\begin{aligned}
\mathbb{E}_{f_1 \dots f_n}\left[F(w_t^k) - f_t^k(w_t^k)\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_1 \dots f_n}\left[f(w_t^k; i) - f(w_t^{k,(i)}; i)\right] \\
&\leq \max_{i \in [n]} G \left\| w_t^k - w_t^{k,(i)} \right\| \\
&\leq \epsilon_{\text{stab}}^{\text{SGD}}(n(k-1) + t - 1; 2k),
\end{aligned}$$

which completes the proof. □

We will also make use of a generalization of Lemma 11 given in [7, Lemma 3.1]. The next lemma is a direct implication of it.

Lemma 9. *The generalized uniform-argument-stability (see Eq. (41)) rate of SGD with step-size $\eta > 0$ for G-Lipschitz convex functions satisfies*

$$\epsilon_{\text{stab}}^{\text{SGD}}(\tau; J) \leq 2G\eta\sqrt{\tau} + 4\eta G J.$$

We are now ready to prove the single-shuffle convergence upper bound.

Proof of Theorem 6 (single-shuffle case). Similarly to the multi-shuffle case, we have;

$$\begin{aligned}
F(\widehat{w}) - F(w^\star) &\leq \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} F(w_t^k) - F(w^\star) \\
&= \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} F(w_t^k) - f_t^k(w^\star) \\
&= \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} F(w_t^k) - f_t^k(w_t^k) + \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} f_t^k(w_t^k) - f_t^k(w^\star) \\
&\leq \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n} F(w_t^k) - f_t^k(w_t^k) + \frac{D^2}{2\eta n K} + \frac{\eta G^2}{2},
\end{aligned}$$

with the last inequality following from the standard nK round regret bound for gradient descent [see e.g., 15]. To bound the other term, we now apply Lemma 8 and Lemma 9 to obtain;

$$\begin{aligned}
\mathbb{E}\left[F(w_t^k) - f_t^k(w_t^k))\right] &\leq G\epsilon_{\text{stab}}^{\text{SGD}}(nk + t; 2k) \\
&\leq 2\eta G^2(\sqrt{n(k-1) + t} + 4k) \\
&\leq 2\eta G^2(\sqrt{nK} + 4K).
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[F(\widehat{w}) - F(w^\star)\right] &\leq \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n}\mathbb{E}\left[F(w_t^k) - f_t^k(w_t^k)\right] + \frac{D^2}{2\eta nK} + \frac{\eta G^2}{2} \\
&\leq \frac{1}{nK}\sum_{k=1}^{K}\sum_{t=1}^{n}(2\eta G^2(\sqrt{nK} + 4K)) + \frac{D^2}{2\eta nK} + \frac{\eta G^2}{2} \\
&\leq 8\eta G^2(\sqrt{nK} + K) + \frac{D^2}{2\eta nK} + \frac{\eta G^2}{2} \\
&\leq \frac{6GD}{n^{1/4}K^{1/4}} + \frac{4K^{1/4}}{n^{3/4}},
\end{aligned}$$

where the last inequality follows from a choice of $\eta = D/(2Gn^{3/4}K^{3/4})$. When $n \geq K$, the above implies

$$\mathbb{E}\left[F(\widehat{w}) - F(w^\star)\right] \leq \frac{10GD}{n^{1/4}K^{1/4}},$$

and concludes the proof. □

E Stability Lemmas

In this section, we provide statements and proofs for several known results relating to stability properties of SGD. For convenience, we repeat the definition of UAS Eq. (4) with notation suitable for SGD;

$$\epsilon_{\text{stab}}^{\text{SGD}}(t) := \max_{f_1,\ldots,f_t,f'}\max_{i\in[t]}\left\|w_{t+1} - w_{t+1}^{(i)}\right\|, \tag{42}$$

where $f_1, \ldots, f_t, f'$ are any sequence of convex Lipschitz losses, w_{t+1} the iterate produced by gradient descent from $w_1 \in W$ on $f_1, \ldots, f_t$, and $w_{t+1}^{(i)}$ the iterate produced from w_1 on the same sequence after replacing f_i with f'.

The next lemma relates the difference between the without-replacement loss distribution and the full batch objective to the uniform stability rate Eq. (42) of the optimization algorithm in question. For a proof see [34] (where it was originally stated for average stability, which is a weaker notion and thus implies the uniform stability case as well).

Lemma 10. *Let $\{f(w;t)\}_{t=1}^n$ be a set of n, G-Lipschitz losses, and $F(w) = \frac{1}{n}\sum_{t=1}^n f(w;t)$. Then, for a uniformly random permutation $\pi\colon [n] \leftrightarrow [n]$, and w_1 independent of π, it holds that*

$$\mathbb{E}_\pi\left[F(w_t) - f(w_t;\pi(t))\right] \leq \frac{(t-1)G}{n}\epsilon_{\text{stab}}^{\text{SGD}}(t-1),$$

where $\epsilon_{\text{stab}}^{\text{SGD}}$ is the stability rate of SGD defined in Eq. (42), and w_t the output of SGD on $\{f(w;s)\}_{s=1}^{t-1}$.

Following are two lemmas providing uniform stability upper bounds for SGD.

Lemma 11. *The uniform argument stability of SGD with step size $\eta > 0$ on convex G-Lipschitz losses is bounded as;*

$$\epsilon_{\text{stab}}^{\text{SGD}}(t) \leq 2G\eta\sqrt{t}.$$

For proof of the above lemma, see [7].

Next, we have standard lemmas providing stability of ERM and regularized ERM in respectively strongly convex and general convex problems.

Lemma 12 (Strongly Convex ERM Stability)**.** *Let $f\colon W \times Z \to \mathbb{R}$ be λ-strongly convex and G-Lipschitz for all $z \in Z$. Then*

$$\left|\mathbb{E}_{S\sim\mathcal{Z}^n}\left[F(w_S^\star) - \widehat{F}(w_S^\star)\right]\right| \leq \frac{G^2}{\lambda n}$$

Proof. Let $\widehat{w}_S := w_S^\star$ denote the empirical risk minimizer, and $\widehat{w}_{S^i}$ the ERM for the training set with the i'th index swapped with a fresh sample z_i'. We have

$$\left|\mathbb{E}_{S\sim\mathcal{Z}^n}\left[F(\widehat{w}_S) - \widehat{F}(\widehat{w}_S)\right]\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[f(\widehat{w}_S; z_i') - f(\widehat{w}_{S^i}; z_i')\right]\right|$$
$$\leq \frac{G}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\widehat{w}_S - \widehat{w}_{S^i}\right\| \leq \frac{4G^2}{\lambda n},$$

where the first inequality is the generalization equals average stability (see e.g., [31]), and the last inequality follows since $\widehat{w}_S$ and $\widehat{w}_{S^i}$ minimize $(1/\lambda n)$-objectives that differ in a $2G$-Lipschitz term. □

Lemma 13 (Regularized ERM Stability)**.** *Let $f\colon W \times Z \to \mathbb{R}$ be G-Lipschitz for all $z \in Z$, and denote the regularized empirical risk minimizer by $\widehat{w}_S^\lambda := \arg\min_{w\in W}\left\{\widehat{F}(w) + \frac{\lambda}{2}\|w\|^2\right\}$. Then*

$$\left|\mathbb{E}_{S\sim\mathcal{Z}^n}\left[F(\widehat{w}_S^\lambda) - \widehat{F}(\widehat{w}_S^\lambda)\right]\right| \leq \frac{G^2}{\lambda n}$$

Proof. Let $F^\lambda(w) := F(w) + \frac{\lambda}{2}\|w\|^2$ and define the regularized empirical loss $\widehat{F}^\lambda$ accordingly. Then we have a λ-strongly convex problem and by Lemma 12,

$$\left|\mathbb{E}_{S\sim\mathcal{Z}^n}\left[F(\widehat{w}_S^\lambda) - \widehat{F}(\widehat{w}_S^\lambda)\right]\right| = \left|\mathbb{E}_{S\sim\mathcal{Z}^n}\left[F^\lambda(\widehat{w}_S^\lambda) - \widehat{F}^\lambda(\widehat{w}_S^\lambda)\right]\right| \leq \frac{4G^2}{\lambda n}.$$

□

F Auxiliary Lemmas

The following provides standard step size dependent lower bounds for convex optimization. See also [1] where similar claims are made in their Lemma 6.2 and implicit in the proof of their Theorem 6.1.

Lemma 14. *For any step-size $\eta > 0$, $T \in \mathbb{N}$ and $d := \lceil 16\eta^2T^2\rceil$, there exists a convex optimization problem $h\colon W \to \mathbb{R}$ where $W \subseteq \mathbb{R}^{d+1}$ is of constant diameter such that*

$$h(\widehat{w}) - \min_{w\in W} h(w) \geq \frac{1}{8}\min\left\{\frac{1}{\eta T} + \eta, 1\right\},$$

and $\widehat{w}$ is any suffix average of T gradient descent step iterates.

Proof. We shall concatenate two objectives; the first is single dimensional and will contribute the η term, the second is d dimensional and will contribute the $1/\eta T$ term.

First objective. Set $f_1(w) := |w - \eta/4|$. Since we initialize at 0, the iterates will "zig-zag" between 0 and $-\eta$. Clearly, any average of iterates is at best $\eta/4$ away from zero loss.

Second objective. Set $f(w) := \max_{i\in[d]}\{w(i)\}$, and note $w^\star = -\frac{1}{\sqrt{d}}\mathbf{1}$ where $\mathbf{1}$ denotes the all ones vector $\in \mathbb{R}^d$. We initialize SGD at $w_1 = 0 \in \mathbb{R}^d$, and follow the gradient steps $\nabla f(w_t) = e_i$ where $i \in [d]$ is one of the coordinates that satisfy $w_t(i) \geq w_t(j)\ \forall j \in [d]$. Hence, for any $t \in [T]$,

$$\|w_{t+1}\|_1 \leq \|w_t\|_1 + \eta\|\nabla f(w_t)\|_1 = \|w_t\|_1 + \eta \leq \cdots \leq \eta t \leq \eta T.$$

By the pigeonhole principle, this implies there must exist some coordinate i such that $w_{T+1}(i) \geq -\eta T/d$. In addition, for any i, $\overline{w}_{\tau:T}(i) \geq w_{T+1}(i)$. Therefore, assuming $8\eta^2T^2 \geq 1$ we conclude;

$$f(\overline{w}_{\tau:T}) - f(w^\star) \geq -\frac{\eta T}{d} + \frac{1}{\sqrt{d}} \geq -\frac{\eta T}{8\eta^2T^2} + \frac{1}{4\eta T} = \frac{1}{8\eta T}.$$

In the case where $8\eta^2T^2 < 1$,

$$f(\overline{w}_{\tau:T}) - f(w^\star) \geq -\eta T + \frac{1}{2} \geq \frac{1}{2} - \frac{1}{2\sqrt{2}} \geq \frac{1}{4},$$

and the result follows. □

Lemma 15. *Let $\{x_1, \ldots, x_n\}$ be a set of real numbers. Then*

$$\frac{1}{n}\sum_{s=1}^{n}\sqrt{\sum_{t=s+1}^{n} x_t^2} \geq \frac{1}{5\sqrt{n}}\sum_{t=n/4}^{n}|x_t|\,.$$

Proof.

$$\begin{aligned}
\frac{1}{n}\sum_{s=1}^{n}\sqrt{\sum_{t=s+1}^{n} x_t^2} &\geq \frac{1}{n}\sum_{s=1}^{n}\frac{1}{\sqrt{n-s}}\sum_{t=s+1}^{n}|x_t| && \text{(by Jensen's inequality)}\\
&= \frac{1}{n}\sum_{t=1}^{n}|x_t|\sum_{s<t}\frac{1}{\sqrt{n-s}}\\
&= \frac{1}{n}\sum_{t=1}^{n}|x_t|\sum_{j=n-t+1}^{n-1}\frac{1}{\sqrt{j}}\\
&\geq \frac{1}{n}\sum_{t=1}^{n}|x_t|\int_{n-t+1}^{n}\frac{1}{\sqrt{x}}\mathrm{d}x\\
&= \frac{2}{n}\sum_{t=1}^{n}|x_t|\left(\sqrt{n}-\sqrt{n-t+1}\right)\\
&\geq \frac{2}{n}\sum_{t=n/4}^{n}|x_t|\left(\sqrt{n}-\sqrt{3n/4}\right)\\
&\geq \frac{2}{\sqrt{n}}\sum_{t=n/4}^{n}|x_t|\left(1-\sqrt{3/4}\right)\\
&\geq \frac{1}{5\sqrt{n}}\sum_{t=n/4}^{n}|x_t|\,.
\end{aligned}$$

□