Generalization Lower Bounds for GD and SGD in Smooth Stochastic Convex Optimization

Peiyuan Zhang
Yale University

Jiaye Teng
Tsinghua University

Jingzhao Zhang
Tsinghua University

Abstract

This work studies the generalization error of gradient methods. More specifically, we focus on how training steps T and step-size η might affect generalization in *smooth* stochastic convex optimization (SCO) problems. Recent works show that in some cases longer training can hurt generalization. Our work reexamines this for smooth SCO and finds that the conclusion can be case-dependent. In particular, we first study SCO problems when the loss is *realizable*, i.e. an optimal solution minimizes all the data points. Our work provides excess risk *lower bounds* for Gradient Descent (GD) and Stochastic Gradient Descent (SGD) and finds that longer training may not hurt generalization. In the short training scenario $\eta T = O(n)$ (n is the sample size), our lower bounds tightly match and certify the respective upper bounds. However, for the long training scenario where $\eta T = \Omega(n)$, our analysis reveals a gap between the lower and upper bounds, indicating that longer training does hurt generalization for realizable objectives. A conjecture is proposed that the gap can be closed by improving upper bounds, supported by analyses in two special instances. Moreover, besides the realizable setup, we also provide first tight excess risk lower bounds for GD and SGD under the general *non-realizable* smooth SCO setting, suggesting that existing stability analyses are tight in step-size and iteration dependence and that overfitting provably happens when there is no interpolating minimum.

1 Introduction

Gradient methods are the predominant algorithms for training neural networks. These methods are not only efficient in time and space, but more importantly, produce solutions that generalize well (He et al., 2016; Vaswani et al., 2017). Understanding why neural networks trained with gradient methods perform well on test data can be challenging, as it results from an interplay between the network architecture, the data distribution, and the training algorithm (Jiang et al., 2019; Zhang et al., 2021). In this work, we aim to study the role of (stochastic) gradient descent and take a step by considering generalization in smooth convex problems.

Much work has been done to analyze gradient descent in convex learning problems. The early approach exploits the convex structure and shows that gradient methods find approximate empirical risk minimizers. Then the generalization can be bounded by uniform convergence (Shalev-Shwartz and Ben-David, 2014). However, this approach is limited in scalability to high dimensions and can be provably vacuous even for the simple task of linear regression (Shalev-Shwartz et al., 2010; Feldman, 2016). An alternative approach (Nemirovskij and Yudin, 1983) that addresses the high-dimension problem is via online-to-batch conversion. This approach achieves minimax sample complexity, but it only applies to single-pass training, whereas in practice, models trained for longer periods can generalize better (Hoffer et al., 2017).

Several recent explanations have been proposed to bridge the gap and bound generalization in multi-pass settings (Soudry et al., 2018; Ji and Telgarsky, 2019; Lyu et al., 2021; Bartlett et al., 2020). These works demonstrate that gradient descent benefits from implicit bias and finds max-margin solutions for classification problems, as well as min-norm solutions for regression problems. However, characterizing the implicit bias for other more general models remains a challenging task. One method that generalizes to a broader range of loss functions and models is the sta-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

bility argument (Bousquet and Elisseeff, 2002; Hardt et al., 2016). This argument shows that if the model and the training method are not overly sensitive to data perturbations, the generalization error can be effectively bounded. However, this argument suffers from a large number of training updates, especially when the step-size is sufficiently large, while in practice, generalization often benefits from longer training time. *It is unclear whether longer training truly hurts generalization in smooth convex learning problems, because the tightness of the growing upper bounds remains unknown, and might just result from an artifact of the analysis.*

In this work, we show that the answer can be case-dependent by analyzing the above problem in smooth *stochastic convex optimization* (SCO). More specifically, we focus on how the *training horizon* ηT might affect the generalization property: the product of the step size η and the number of iterations T is a better measure for training intensity as a large number of iterations may not even train a model if η is close to zero. While several recent works have established fast convergence rates in test error when ηT is not too large under the *realizable* condition (Lei and Ying, 2020; Nikolakakis et al., 2022; Schliserman and Koren, 2022), our work provides the first tight lower bounds in these scenarios and suggests that known bounds are tight in some settings but likely not when ηT is large.

Our contributions. Let η represent the step-size in gradient methods, T denote the iteration number, and n denote the sample size. Our contributions are as follows and presented in Table 1:

- For realizable SCO, we notice a gap between two types of analysis, as shown in Table 1:
  1. when $\eta T = O(n)$, we prove matching lower bounds for the excess population risk for GD and SGD under the smooth and realizable SCO setting;
  2. when $\eta T = \Omega(n)$, we provide a lower bound construction that suggests a gap exists between upper and lower bounds, and that longer training may not hurt generalization in this scenario.
- We then provide a tight lower bound $\Omega\left(\frac{1}{\eta T} + \frac{\eta T}{n}\right)$ for the smooth non-realizable SCO.
- We conjecture that the upper bound for realizable SCO when $\eta T = \Omega(n)$ is not tight. We provide evidence for the conjecture in two special scenarios: (1) one-dimensional convex problems and (2) high-dimensional linear regression.

Our results are first to provide generalization lower bounds for the smooth and realizable SCO setting and are based on novel constructions. In this regard, we first reveal that that the *training horizon* ηT is crucial in deciding overfitting and $1/n$ sample rate. These lower bounds entailed in Table 1, offer insights and answers to the question of how longer training impacts generalization error. For non-realizable cases, our lower bound suggests that training for a longer time can provably lead to overfitting, even for smooth convex problems. For realizable cases, our lower bounds suggest that longer training might reduce the generalization error. Moreover, our new lower bounds in the smooth setting, compared with those known in the nonsmooth setting, suggest that smoothness and realizability together might explain why training longer does not lead to overfitting.

Before we end this discussion, we note that the open question *whether the upper bound in long horizon setup is tight* only arises due to our analyses identifying the role of ηT in the lower bound. An improved upper bound result would be a natural and challenging research problem motivated by this work.

Notations. For any positive integer n, we denote the set $[n] := \{1, 2, \ldots, n\}$. $\|\cdot\|$ denotes the l_2 norm for vectors. We use $\mathrm{Bern}(p)$ to denote the Bernoulli distribution with probability p to be 1 and $\mathrm{Unif}(S)$ to denote the uniform distribution over set S. Occasionally we will use capital letters, i.e. W, to denote "large" vectors, in contrast to usual vectors like v, w. Also, let $\mathbf{1}$ be an all-one vector and $\boldsymbol{e}_i$ be the vector with a 1 in the i-th coordinate and 0 elsewhere.

2 Preliminaries

Following Amir et al. (2021b) Bassily et al. (2020) and Shalev-Shwartz et al. (2009), we study the generalization error of *stochastic convex optimization (SCO)* problem. In SCO, we receive a dataset of finite samples $S = \{z_1, \ldots, z_n\}$, where each z_i is i.i.d. drawn from an unknown distribution D over sample space $\mathcal{Z}$ and n is the size of the dataset. Our goal is to find a model parameterized by $w \in \mathcal{W} \subseteq \mathbb{R}^d$ that minimizes the *population* (or expected) risk over D, defined as:

$$F(w) = \mathbb{E}_{z\sim D}[f(w, z)] \tag{1}$$

where $f(w, z) : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is the loss function evaluated on a single example $z \in \mathcal{Z}$.

Since the population loss F is typically inaccessible, we instead employ an averaged substitute on sample S, known as the *empirical* risk:

$$F_S(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, z_i), \qquad S \sim D^n. \tag{2}$$

		GD	SGD	Best Sample Complexity	Overfitting for large ηT
Realizable	Upper bound	$O\left(\frac{1}{\eta T}+\frac{1}{n}+\frac{\eta T}{n^2}\right)$ Nikolakakis et al. (2022)†	$O\left(\frac{1}{\eta T}+\frac{\eta}{n}+\frac{\eta T}{n^2}\right)$ Lei and Ying (2020)	$O\left(1/n\right)$‡	Yes
	Lower bound $(\eta T = O(n))$	$\Omega\left(\frac{1}{\eta T}+\frac{1}{n}+\frac{\eta T}{n^2}\right)$ (Theorem. 3.1)	$\Omega\left(\frac{1}{\eta T}+\frac{1}{n}+\frac{\eta T}{n^2}\right)$ (Theorem. 3.1)	$\Omega\left(1/n\right)$	N.A.
	Lower bound $(\eta T = \Omega(n))$	$\Omega\left(\frac{1}{\eta T}+\frac{1}{n}\right)$ (Theorem. 3.2)	$\Omega\left(\frac{1}{\eta T}+\frac{1}{n}\right)$ (Theorem. 3.2)	$\Omega\left(1/n\right)$	No
Non-realizable	Upper bound	$O\left(\frac{1}{\eta T}+\frac{\eta T}{n}\right)$ Hardt et al. (2016)	$O\left(\frac{1}{\eta T}+\frac{\eta T}{n}\right)$ Hardt et al. (2016)	$O\left(1/\sqrt{n}\right)$	Yes
	Lower bound	$\Omega\left(\frac{1}{\eta T}+\frac{\eta T}{n}\right)$ (Theorem. 3.3)	$\Omega\left(\frac{1}{\eta T}+\frac{\eta T}{n}\right)$ (Theorem. 3.4)	$\Omega\left(1/\sqrt{n}\right)$	Yes

Table 1: Summary of our results. We present our lower bounds and compare them with existing upper bounds. In particular, we split into three settings: non-realizable, realizable under $\eta T = O(n)$ and realizable under $\eta T = \Omega(n)$. For each setting, we provide lower bounds for the excess risk of GD and SGD. We also provide the best possible sample complexity and whether overfitting happens when ηT is large.

† The bound in η, T and n is not explicitly stated in Nikolakakis et al. (2022). For the expression, please refer to a derivation in Appendix D.3.

‡ The best sample complexity $O(1/n)$ is **only** possible when $\eta T = O(n)$, and can be arbitrarily bad when $\eta T = \Omega(n)$. See discussion in Section 3.

Given dataset S and any (stochastic) algorithm $\mathcal{A}$, we denote $\mathcal{A}[S]$ as the output of running $\mathcal{A}$ on the training sample S. In this paper, we are interested in bounding the *excess population risk* of $\mathcal{A}[S]$:

$$\mathbb{E}_{S,\mathcal{A}}[F(\mathcal{A}[S])] - \min_{w\in\mathcal{W}} F(w),$$

where the expectation is taken over the randomness of sample S and algorithm $\mathcal{A}$.

2.1 Gradient Methods

In this work, we focus on understanding the excess risk for two simplified algorithms: Gradient Descent (GD) and Stochastic Gradient Descent (SGD). Gradient descent is one of the most well-known optimization methods. At iteration t, GD employs the following recurrence:

$$w_{t+1} = w_t - \eta \nabla F_S(w_t), \tag{3}$$

where $\eta > 0$ is the step-size and $\nabla F_S(w)$ is the average stochastic gradient on sample set S. We usually employ the time average $w_{\mathrm{GD}} := \bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ as the output of GD.

In practice, many practitioners favor the Stochastic Gradient Descent (SGD) method over GD for its computational efficiency. In this work, we study standard SGD, i.e., in iteration $t \in [T]$,

$$w_{t+1} = w_t - \eta \nabla f(w_t, z_{i_t}), \tag{4}$$

where z_{i_t} is uniformly sampled from S *with replacement* as $i_t \sim \mathrm{Unif}([n])$. The output for SGD is the average $w_{\mathrm{SGD}} := \bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$.

2.2 Smooth Stochastic Convex Optimization

In order to derive non-vacuous bounds on the excess risk for SCO, we make assumptions on the properties of $f(w,z)$. First, we assume the access to the value $f(w,z)$ and the unbiased stochastic gradient estimator $\nabla f(w,z)$ for any $w \in \mathcal{W}$ and $z \in \mathcal{Z}$.

When the function is nonsmooth, this problem has been extensively studied (Bassily et al., 2020), and known rates were proven to be optimal (Amir et al., 2021b; Sekhari et al., 2021; Nemirovskij and Yudin, 1983). However, less is known when the function is differentiable and smooth. Indeed, while upper bounds have been well-established in literature (Hardt et al., 2016; Lei and Ying, 2020; Nikolakakis et al., 2022), the optimality of these results is yet to be certified by corresponding lower bounds. In this work, we aim to provide lower bounds for the smooth SCO setting and make the following assumptions.

Definition 2.1. *$f(w,z)$ is L-smooth if it satisfies $\|\nabla f(w_1,z)-\nabla f(w_2,z)\| \le L\|w_1-w_2\|$ for any w_1, w_2 and $z \in \mathcal{Z}$.*

Definition 2.2. *$f(w,z)$ is convex if it satisfies $f(w_1,z) \ge f(w_2,z) + \langle w_1 - w_2, \nabla f(w_2,z)\rangle$ for any w_1, w_2 and $z \in \mathcal{Z}$.*

Realizable smooth SCO. The smooth SCO problem can be divided into two cases depending on whether the optimal solution minimizes all data points simultaneously. When this happens, it is usually referred to as a *realizable* setting, formally defined below.

Definition 2.3. *We say that $f(w,z)$, $z \in \mathcal{Z}$ formalizes a realizable setting if for any $z \in \mathcal{Z}$*

$$f(w^*,z) = \min_w f(w,z),$$

where $w^* = \arg\min_w F(w)$.

If $f(w,z)$ is smooth, convex, and realizable, we refer to the setting as the *realizable* smooth SCO. The realizable condition implies immediately the property called *weak growth condition*, which is stated in the following lemma.

Lemma 2.4. *If $f(w,z), z \in \mathcal{Z}$ is realizable and L-smooth, then for any $w, z \in \mathcal{Z}$ it holds*

$$\|\nabla f(w,z)\|^2 \le 2L\left(f(w,z) - f(w^*,z)\right).$$

The growth condition connects the rates at which the stochastic gradients **shrink** relative to its value. It is widely employed in stochastic optimization literature to improve the convergence rate of SGD and GD under overparameterized or realizable settings (Vaswani et al., 2019). Recent papers (Lei and Ying, 2020; Schliserman and Koren, 2022; Nikolakakis et al., 2022) focus on the generalization bound under realizable smooth SCO and also suggest that such an assumption improves the sample complexity upper bounds.

Non-realizable smooth SCO. We say a convex learning problem is non-realizable if it does not satisfy the condition in Definition 2.3. In this setting, known upper bounds for sample complexity actually yield a slower convergence rate at $O\left(1/\sqrt{n}\right)$ (Hardt et al., 2016) when we set $\eta T = \Theta(\sqrt{n})$, which suggests that longer training leads to overfitting.

3 Main Theorems: Lower Bounds in Smooth SCO

In this section, we will present and discuss our main result on generalization lower bounds under the smooth SCO setting. The lower bounds are reported and compared with the corresponding upper bounds in Table 1.

3.1 Realizable Setting

The necessity of lower bound in the realizable smooth SCO setting arises from the theory-practice gap on how longer training might affect generalization. To address this, we introduce the *training horizon* ηT, as the product of the step-size η and the number of iterations T. This quantity better measures the training process's intensity compared to relying solely on T. This is because an arbitrarily small step-size with a large T does not necessarily decrease the optimization error to convergence. Furthermore, ηT marks a critical phase transition in existing upper bound results: (Lei and Ying, 2020) establishes the following bound for excess population risk of SGD

$$\mathbb{E}[F(w_{\mathrm{SGD}})] - \min_{w\in\mathcal{W}} F(w) = O\left(\frac{1}{\eta T} + \frac{\eta}{n} + \frac{\eta T}{n^2}\right). \quad (5)$$

The bound implies generalization property is decided by ηT. More precisely, an improved sample rate $O(1/n)$ is attained when the training horizon satisfies $\eta T = O(n)$. *Otherwise, if ηT is sufficiently large, say $\eta T = n^2$, the generalization error escalates to $O(1)$ and overfits. This contradicts the empirical observation that longer training helps generalization.* Similar observation also applies to the analysis of GD.

Challenges in Lower Bound Construction. To bridge the theory-practice gap, we analyze the relationship between overfitting, generalization error, and training horizon ηT from a *lower bound* perspective. The challenging task necessitates novel hard constructions for the excess population risk, which we elaborate on below.

Existing generalization lower bounds (Amir et al., 2021b,a; Shalev-Shwartz et al., 2009) primarily address the **nonsmooth or the non-realizable** setting. In this scenario, hard constructions are devised by leveraging non-continuity of subgradient and non-shrinking noise[1], allowing the stochastic descent $-\nabla f(w,z)$ to admit an opposite direction from the minimum of the population risk w^*,

$$\langle -\nabla f(w,z), w - w^*\rangle > 0.$$

[1] Under the realizable setting, the stochastic gradient shrinks as in the growth condition, Lemma 2.4, whereas in the non-realizable setting, this does not necessarily holds. We refer to this as the *non-shrinking* property.

In contrast, the above situation can **never** occur in the **realizable and smooth** SCO. Realizable SCO is characterized by continuous gradients and an interpolating global minimum, where we have

$$\langle -\nabla f(w,z), w - w^* \rangle \leq 0.$$

The property imposes more difficulty in keeping the gradient steps away from the minimum. We overcome this challenge by exploiting stochasticity in the high dimensional regime where the dimension grows exponentially in the sample size $d = \Theta(e^n)$. Through this approach, we ensure that, at each iteration and with high probability, multiple coordinates are pushed away from the minimum by coordinate-specific stochastic gradients. Combined with other building blocks, this leads to a large gap in the excess risk. A sketch proof can be found in Section 5 and the complete proof is provided in the Appendix. Our constructions are novel and therefore are not comparable to the aforementioned lower bound constructions in the nonsmooth SCO setting. We will now present our main results and discuss the long and short-time horizon (defined by ηT) regimes separately below.

The Short Horizon Regime $\eta T = O(n)$. We first provide our results for the realizable setting, when condition $\eta T = O(n)$ is satisfied. The next theorem characterizes the lower bounds for GD and SGD.

Theorem 3.1. *For every $\eta > 0$, $T > 1$ with $1/T \leq \eta = O(1)^2$, if condition $T = \mathcal{O}(n)$ holds, then there exists a distribution D, and a convex, 1-smooth and realizable $f(w,z) : \mathbb{R}^d \to \mathbb{R}$ for every $z \sim D$, such that, with a bounded initialization $\|w_1 - w^*\| = O(1)$, the output w_{GD} for GD satisfies*

$$\mathbb{E}[F(w_{GD})] - F(w^*) = \Omega\left(\frac{1}{\eta T} + \frac{1}{n} + \frac{\eta T}{n^2}\right).$$

Similarly, the output w_{SGD} for SGD satisfies

$$\mathbb{E}[F(w_{SGD})] - F(w^*) = \Omega\left(\frac{1}{\eta T} + \frac{1}{n} + \frac{\eta T}{n^2}\right).$$

The excess risk is composite and consists of three individual terms, which we comment on below:

- The term, $\Omega\left(\eta T/n^2\right)$, represents the generalization error and comes from our major new construction, and is formally defined in Section 5 and Appendix A.3. This construction resolves the challenge discussed earlier;
- The term $\Omega\left(1/\eta T\right)$ reflects the optimization error, which is redesigned to suit the realizable setting;
- The term $\Omega\left(1/n\right)$ comes from a universal hard instance that holds for any deterministic or stochastic gradient methods.

Notice that the term $\Omega\left(1/n\right)$ does not suggest the rest two terms are vacuous since they are hard in the sense of characterizing the relationship between η, T and n. The proof for each term is provided in Lemma D.2, Appendix D.2 and Lemma D.1, Appendix D.1.

Theorem 3.1 suggests that known upper bounds are tight not only in sample rate but also in T and η dependence. More specifically, the lower bound for GD tightly matches the upper bound in Nikolakakis et al. (2022), and the lower bound for SGD almost tightly matches the lower bound in Lei and Ying (2020) up to a η factor in the second term. Please refer to Table 1 for a comparison. We will combine the discussion for GD and SGD due to their similarity. Both upper and lower bounds are non-vacuous only when $T = \Theta(n)$ and $\eta = \Theta(1)$: under this configuration, we obtain the optimal sample rate lower bound $\Omega\left(1/n\right)$ from Theorem 3.1, which matches the sample rate upper bound $O\left(1/n\right)$ under the regime of $\eta T = O\left(n\right)$. We will see in the next subsection that the conclusion is different when $\eta T = \Omega\left(n\right)$.

The Long Horizon Regime $\eta T = \Omega\left(n\right)$. We turn to the large or infinite training horizon setting, i.e. $\eta T = \Omega\left(n\right)$. The aforementioned upper bounds suggest overfitting will happen under the regime, contradicting common observations. We use our lower bound results, presented in the following theorem, to consolidate that longer training does not lead to overfitting.

Theorem 3.2. *For every $\eta > 0$, $T > 1$, if condition $\eta T = \Omega\left(n\right)$ holds, then there exists a distribution D, and a convex, 1-smooth and realizable $f(w,z) : \mathbb{R}^d \to \mathbb{R}$ for every $z \sim D$ such that, with a bounded initialization $\|w_1 - w^*\| = O\left(1\right)$, the output w_{GD} for GD satisfies*

$$\mathbb{E}[F(w_{GD})] - F(w^*) = \Omega\left(\frac{1}{\eta T} + \frac{1}{n}\right).$$

Similarly, the output w_{SGD} for SGD satisfies

$$\mathbb{E}[F(w_{SGD})] - F(w^*) = \Omega\left(\frac{1}{\eta T} + \frac{1}{n}\right).$$

In Theorem 3.2, GD and SGD have the same upper and lower bounds for excess population risk. Theorem 3.2 indicates that, different from the case $\eta T = O\left(n\right)$, our lower bound for both GD and SGD **does not** match the corresponding upper bounds in Lei and Ying (2020); Nikolakakis et al. (2022) (see Table 1). *For lower bound,*

[2] This is a mild condition because (1) step-size cannot exceed $O\left(1\right)$, in order to make the optimization method converge for $O\left(1\right)$-smooth function, (2) an overly small step-size η cannot even guarantee the convergence in the optimization sense and T is arbitrarily large to ensure $\eta T \geq 1$. We will assume this holds in the statement of rest theorems and lemmas.

we achieve best sample rate $\Omega(1/n)$ *as long as* $\eta T \geq n$*, whereas for upper bound, the best sample rate* $O(1/n)$ *is obtained* only *when we set* $\eta T = n$. To conclude, while the upper bound indicates longer training leads to overfitting, the lower bound suggests the opposite under the realizable setting.

To establish Theorem 3.2, we employ a strategy similar to the proof of Theorem 3.1. Especially for the generalization term, albeit the similar construction with Theorem 3.1, the **difference** comes from lower bounding $1-(1-\eta/n)^T$ in two regimes: when $\eta T = O(n)$, we have $1-(1-\eta/n)^T = \Omega(1)$ and when $\eta T = \Omega(n)$, we have $1-(1-\eta/n)^T = \Omega(\eta T/n)$. This then leads to a difference in the absolute value of each coordinate. The details are postponed to Appendix A.2.

Theorem 3.1 and Theorem 3.2 give a complete description of excess population risk and are the first to establish lower bounds under the smooth and realizable setting. The novel results support the observation that longer training does not hurt generalization and reveals a gap between lower and upper bounds. We conjecture the sample rate bound under a large or infinite time horizon can be closed by proving upper bound $O(1/n)$ is achievable for GD even when ηT goes to infinity. We will discuss the conjecture and provide several evidences in Section 4.

3.2 Non-realizable Setting

In this subsection, we also provide novel results on the lower bounds of both GD and SGD in the *non-realizable* smooth SCO setting, which to the best of our knowledge have not been previously reported. The lower bound for the excess risk of GD is stated in the following theorem.

Theorem 3.3. *For any* $\eta > 0$*,* $T > 1$*, there exists a convex,* 1*-smooth* $f(w,z) : \mathbb{R} \to \mathbb{R}$ *for every* $z \in \mathcal{Z}$*, and a distribution* D *such that, with a bounded initialization* $\|w_1 - w^*\| = O(1)$*, the output* w_{GD} *for GD satisfies*

$$\mathbb{E}[F(w_T)] - F(w^*) = \Omega\left(\frac{1}{\eta T} + \frac{\eta T}{n}\right).$$

The lower bound in Theorem 3.3 tightly matches the corresponding upper bound established in Hardt et al. (2016) (see Table 1). It can be translated to a lower bound of sample rate: for any $T > 1$, by setting $\eta = \sqrt{n}/T$, we derive a $\Omega(1/\sqrt{n})$ bound which certifies the optimality of the existing upper bound.

To the best of our knowledge, this is the first such result for GD. A recent work provides a lower bound for the *uniform stability* of (S)GD (Zhang et al., 2022) under smooth SCO, but it does not directly imply a bound on the *excess risk*. More importantly, our construction has low dimension dependence by establishing lower bounds in $d = 1$ (and hence in any $d \geq 1$). As mentioned as an open question in Amir et al. (2021b), improving d dependence from exponential to polynomial in n is of major practicality and significance.

We highlight the key intuitions below. Similar to the realizable setting, the term $\Omega(1/(\eta T))$ reflects the optimization error. The major challenge in the proof of Theorem 3.3 is the term $\Omega(\eta T/m)$. To overcome this difficulty, we employ a novel technique inspired by Theorem 3 and Lemma 7 in Sekhari et al. (2021): in the non-realizable setting, the stochastic gradient does not necessarily scale down with the value of $f(w,z)$. As a result, by utilizing an *anti-concentration* argument, we show that with non-vanishing probability $\Omega(1)$, the absolute value of w_t increases by a rate of $\Omega(\eta/\sqrt{n})$ in each step. Then, the calculation suggests a $\Omega(\eta T/n)$ bound for the function value. This is fundamentally different from the technique in Sekhari et al. (2021), which tackles the nonsmooth setting. The details are in Appendix B.1. In the meanwhile, the term $\Omega(1/\eta T)$ reflects the optimization error and the proof is provided in Lemma D.2, Appendix D.2. A similar result holds for SGD in the following theorem.

Theorem 3.4. *For any* $\eta > 0$*,* $T > 1$*, there exists a convex,* 1*-smooth* $f(w,z) : \mathbb{R} \to \mathbb{R}$ *for every* $z \in \mathcal{Z}$*, and a distribution* D *such that, with a bounded initialization* $\|w_1 - w^*\| = O(1)$*, the output* w_{SGD} *for SGD satisfies*

$$\mathbb{E}[F(w_{SGD})] - F(w^*) = \Omega\left(\frac{1}{\eta T} + \frac{\eta T}{n}\right).$$

This also matches the SGD upper bound in Hardt et al. (2016) and implies a sample rate bound $\Omega(1/\sqrt{n})$ if we set $\eta = \sqrt{n}/T$ for any T. We emphasize the bound for SGD is novel compared with existing works: it is a folklore that in Nemirovskij and Yudin (1983), single-pass SGD ($T = n$) achieves a sample rate lower bounds for Lipschitz convex functions (where a smooth function within a bounded domain is automatically Lipschitz). Yet, our result is the first to provide an explicit dependence on T and η and applies to an arbitrary number of updates. It shows that training longer can provably lead to overfitting, and answers the question raised in the introduction for the non-realizable setting.

4 Upper Bounds in the Long Horizon Regime $\eta T = \Omega(n)$

In Section 3, we establish lower bounds for both realizable and non-realizable cases. For non-realizable losses, both the upper and lower bounds are tight regardless of the relationship between T and n. The result differs for the realizable cases: while upper bound results suffer from large training time, our lower bounds say that overfitting does not happen. It is natural to ask

Can we close the gap between upper and lower bounds for realizable SCO when ηT goes to infinity?

We conjecture that the above problem can be tackled by proving GD and SGD can achieve $O(1/n)$ even when ηT goes to infinity. In the section, we provide evidence to support the conjecture: we consider the examples of one-dimensional function and linear regression. In both examples, $\Theta(1/n)$ sample complexity is achieved for GD and SGD when ηT is large.

4.1 One-dimensional Feasibility

We support our conjecture by providing the first evidence in dimension one: under $d = 1$, we close the gap between the upper and lower bound by establishing $\Theta(1/n)$ sample complexity in the rest part of the subsection. The first result is Lemma 4.1, which establishes an upper bound for SGD based on Lei and Ying (2020).

Lemma 4.1. *In dimension one, if $f(w, z)$ is convex, 1-smooth and realizable with $z \sim D$, then for every $\eta = \Theta(1)$, there exists $T_0 = \Theta(n)$ such that for $T \geq T_0$, the output w_{SGD} of SGD satisfies*

$$\mathbb{E}[F(w_{SGD})] - F(w^*) = O(1/n) .$$

A similar result can be established for GD as below. The proofs are postponed to Appendix C.

Lemma 4.2. *In dimension one, if $f(w, z)$ is convex, 1-smooth and realizable with $z \sim D$, then for every $\eta = \Theta(1)$, there exists $T_0 = \Theta(n)$ such that for $T \geq T_0$, the output w_{GD} of GD satisfies*

$$\mathbb{E}[F(w_{GD})] - F(w^*) = O(1/n) .$$

Unfortunately, we cannot employ the same technique to extend the result to a high-dimensional case. However, we show that the gap can be closed for the special case of linear regression in the high-dimensional regime in the next subsection.

4.2 Linear Regression

In this subsection, we demonstrate that when $\eta T = \Omega(n)$, $\Theta(1/n)$ can be achieved on *linear regression* problem, whether underparameterized ($d < n$) or overparameterized ($d \geq n$). In realizable (or noiseless) linear regression problems, the i-th sample $z_i = (x_i, y_i)$ in dataset $S = \{z_1, \dots, z_n\}$ satisfies that $y_i = x_i^\top w^*$ and x_i is i.i.d. drawn from an unknown distribution. Under the linear predictor $x_i^\top w$, the loss term is defined as $f(w, z) = (y_i - x_i^\top w)^2$. Under this regime, a bounded feature $\|x\| = O(1)$ suffices to guarantee that $f(w, z)$ is convex, $O(1)$-smooth, and realizable. In this case, the upper bound would be $O(1/n)$ and the lower bound would be $\mathcal{O}(\log^3 n/n)$, which is optimal up to a log-factor. We present Lemma 4.3, which establishes an upper bound using local Rademacher complexity.

Lemma 4.3 (From Srebro et al. (2010))**.** *In the realizable linear regression cases, for every $\eta > 0$ and $T > 1$, if the feature x_i is bounded, it holds that for the output of SGD*

$$\mathbb{E}[F(w_{SGD})] - F(w^*) = O\left(1/(\eta T) + \log^3 n/n\right) ,$$

and also the output for GD

$$\mathbb{E}[F(w_{GD})] - F(w^*) = O\left(1/(\eta T) + \log^3 n/n\right) .$$

Proof. One could apply Theorem 1 in Srebro et al. (2010). Specifically, we plug in the realizable assumption and the Rademacher complexity of linear functions, which is in order $O(1/\sqrt{n})$ in bounded norm cases. □

Lemma 4.3 established a sample complexity rate of $\Theta(1/n)$ for linear regression when T grows large. Our evidence on both dimension one case and regression suggests the gap in the regime $\eta T = \Omega(n)$ might be closed by improving the upper bounds of excess risk or sample complexity. However, the approach does not generalize to general convex functions due to as convex functions have much larger Rademacher complexity. We hope our analysis can motivate future exploration into the topic.

5 Proof Overviews

In this section, we provide a brief overview of our technique used in the proofs of theorems for realizable cases in Section 3. In particular, we will focus on the lower bound construction for the output of GD when $\eta T = O(n)$, i.e. first part in Theorem 3.1, to showcase the major intuition and idea behind our constructions.

As discussed in the above section, the main difficulty in the GD part of Theorem 3.1 lies in proving

$$\mathbb{E}[F(w_{\mathrm{GD}})] - F(w^*) = \Omega\left(\eta T/n^2\right) . \tag{6}$$

The proof of rest terms is based on easier constructions and we recommend referring to Lemma D.1 and Lemma D.2 in Appendix D. These lemmas are general and hold for any deterministic or stochastic gradient methods. Here we focus on the proof of (6). Our technique is inspired by the work of Amir et al. (2021b); Sekhari et al. (2021). However, their construction critically relies on nonsmoothness and nonrealizability and therefore does not apply to our setting. The major challenge is that, when the objective is smooth and realizable, the minima of empirical and population risk coincide, therefore the construction in Amir et al. (2021b) does not work. Novel constructions are required for the smooth realizable SCO lower bound, as described below.

We start by considering running GD on the following 2-dimensional quadratic function

$$h(x,y) = \frac{\alpha x^2}{2} + \frac{y^2}{2} - 2\sqrt{\alpha} xy = \frac{1}{2}\left|\sqrt{\alpha}x - y\right|^2$$

with step-size η and initialization $x_1 = 1$ and $y_1 = 0$. We choose a small enough $\alpha = 1/\eta T \ll 1$. In every iteration, since α is small, x is pulled back to zero slowly: it is easy to lower bound the value as

$$\begin{aligned} x_{t+1} &\geq (1-\alpha\eta)x_t \geq \Theta(e^{-\alpha\eta t}) \cdot x_1 \\ &\geq \Theta(e^{-t/T}) \cdot x_1 \geq \Theta(1) \cdot x_1 = \Theta(1). \end{aligned}$$

Hence $x_t = \Omega\left(1\right)$ for any $t \in [T]$. Meanwhile, coordinate y is simultaneously (1) pushed away from zero by x on the scale of $\Omega\left(\eta\sqrt{\alpha}\right)$ and (2) pulled back towards zero by itself. As a result, despite the pulling influence, we can still guarantee that y_t is bounded away from zero for all $t \in [T]$. We now want to improve over the naive two-dimensional quadratic example to make sure that multiple coordinates are bounded away from zero. This intuitively might provide a hard instance for the GD algorithm. Also, we hope stochasticity plays a role in the hard instance such that we can introduce the factor of n. We then devise the following instance $g(w,z) : \mathbb{R}^{n+1} \times \mathcal{Z} \to \mathbb{R}$ belonging to the realizable smooth SCO setting:

$$\begin{aligned} g(w, z=i) &= \tfrac{\alpha}{2}x^2 + \tfrac{1}{2}\big(y(i)\big)^2 - \sqrt{\alpha}x \cdot y(i) \\ &= \tfrac{1}{2}\left|\sqrt{\alpha}x - y(i)\right|^2 \end{aligned}$$

where $w = (x, y)$, $x \in \mathbb{R}$, $y \in \mathbb{R}^n$ and $z \sim \mathrm{Unif}([n])$. We still set parameter α to be $1/(\eta T)$. We are given a dataset S of n examples i.i.d. from the distribution. This leads to population loss

$$G(w) = \mathbb{E}_{z\sim\mathrm{Unif}([n])}[g(w,z)] = \tfrac{1}{2n}\left\|y - \sqrt{\alpha}x\right\|^2.$$

We generalize the idea from $d = 2$ to $d = n + 1$ dimension. To this end, we need every example $z_i \in S$ corresponding to one coordinate $y(i)$. This is, however, an improbable event that occurs with probability $\Theta(\sqrt{n} \cdot e^{-n})$. We use the intuition from Amir et al. (2021b); Sekhari et al. (2021): if we consider multiple independent copies of $g(w, z)$, then with probability $\Theta(1)$, there exists at least one copy that satisfies the condition.

We focus on the particular copy only. Our calculations shows that under assumption $\eta T = O\left(n\right)$, it holds that, for any $t \in [T]$, (1) $x_t = \Theta(1)$ and (2) $y_t(i) = \Omega\left(\sqrt{\eta t}/n\right)$ for any coordinates $i \in [n]$. With a slight abuse of the notation w, we put everything together and guarantee

$$F(w_{\mathrm{GD}}) - F(w^*) = \Omega\left(\|y_T\|^2/(2n)\right) = \Omega\left(\frac{\eta T}{n^2}\right).$$

The details in the proof of Theorem 3.1 can be found in Appendix A.1. The idea behind the proof for the case $\eta T = \Omega\left(n\right)$ (Theorem 3.2) differs only in calculations and hence we omit the repetition. Proof for SGD is also similar. The details of proof for other theorems can be found in Appendix A.

6 Additional Related Work

Generalization in stochastic convex optimization has been extensively explored in the literature (Boyd et al., 2004; Shalev-Shwartz and Ben-David, 2014), with one-pass SGD (Pillaud-Vivien et al., 2018a), multi-pass SGD (Pillaud-Vivien et al., 2018b; Sekhari et al., 2021; Lei et al., 2021), DP-SGD (Bassily et al., 2019; Ma et al., 2022), ERM solution (Feldman, 2016; Aubin et al., 2020) and so forth. One of the most famous results is that one-pass SGD can achieve an optimal error rate of $\mathcal{O}(1/\sqrt{n})$, even in the presence of non-smooth loss (Nemirovskij and Yudin, 1983).

However, for realizable problems, existing analyses typically focus only on upper bounds (Lei and Ying, 2020; Nikolakakis et al., 2022; Schliserman and Koren, 2022; Taheri and Thrampoulidis, 2023) and matching lower bounds are lacking. Realizability is closely related to label noise, which can have a substantial impact on generalization (Song et al., 2019; Harutyunyan et al., 2020; Teng et al., 2022; Wen et al., 2022). For lower bounds, Amir et al. (2021b) show that no less than $\Omega\left(1/\epsilon^4\right)$ steps are needed for GD to achieve ϵ-excess risk, whereas SGD needs only $O\left(1/\epsilon^2\right)$, whereas some other works (including our work) focus on the sample complexity bound. Sekhari et al. (2021) further indicate that GD suffers from a $\Omega\left(1/n^{5/12}\right)$ sample complexity, which is slower than the well-established bound $\Theta(1/\sqrt{n})$ for SGD (Nemirovskij and Yudin, 1983). Another line of lower bounds focuses on the failure of analyses rather than that of the algorithm. For instance, despite the optimal rate of $\mathcal{O}(1/\sqrt{n})$ in convex optimization, uniform convergence only returns a lower bound of $\Omega(\sqrt{d/n})$ (Shalev-Shwartz et al., 2010; Feldman, 2016), leading to a constant lower bound in overparameterized regimes. Other works further illustrate the inherent weakness of uniform convergence (Nagarajan and Kolter, 2019; Glasgow et al., 2022), or of stability-based bounds (Bassily et al., 2020).

To bridge the gap between lower and upper bound, a fast rate upper bound in order $O(1/n)$ is required.

One of the most well-known fast-rate bounds is local Rademacher complexity, which works well under low-noise regimes (Bartlett et al., 2005). However, it typically relies on a specific function class and may not be directly applied to the general convex optimization regimes (Steinwart and Scovel, 2007; Srebro et al., 2010; Zhou et al., 2021). Alternatively, stability-based analyses have shown promise and work well in convex optimization regimes, which have the potential to provide fast-rate generalization bound (Bousquet and Elisseeff, 2002; Hardt et al., 2016; Feldman and Vondrak, 2019; Zhang et al., 2022). In addition to these bounds, one can also derive fast rate bound for finite-dimensional cases (Lee et al., 1996; Bousquet, 2002), aggregation (Tsybakov, 2004; Chesneau and Lecué, 2009; Dalalyan et al., 2018), PAC-Bayesian and information-based analysis (Yang et al., 2019; Grunwald et al., 2021).

7 Conclusion

In this work, we focus on generalization bounds under the smooth SCO setting. In particular, we provide lower bounds for excess risk as a function sample size n, the learning rate η and the iteration T under three settings: (1) non-realizable, (2) realizable with $\eta T = O(n)$, and (3) realizable with $\eta T = \Omega(n)$. For the first two cases, our lower bounds match the corresponding upper bounds and certificate the optimal sample complexity. Nevertheless, under the realizable case with $\eta T = O(n)$, we observe a gap between existing upper bounds and lower bounds. We conjecture that this gap can be closed by improving the upper bound under the long-time horizon regime, and provide evidence in the one-dimensional problem and the linear regression problem to support our hypothesis.

References

Idan Amir, Yair Carmon, Tomer Koren, and Roi Livni. Never go full batch (in stochastic convex optimization). *Advances in Neural Information Processing Systems*, 34:25033–25043, 2021a.

Idan Amir, Tomer Koren, and Roi Livni. Sgd generalizes better than gd (and regularization doesn't help). In *Conference on Learning Theory*, pages 63–92. PMLR, 2021b.

Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

Peter Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

Olivier Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Christophe Chesneau and Guillaume Lecué. Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. *Statistica Sinica*, pages 1407–1417, 2009.

AS Dalalyan, E Grappin, and Q Paris. On the exponentially weighted aggregate with the laplace prior. *Annals of Statistics*, 46(5):2452–2478, 2018.

Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. *Advances in Neural Information Processing Systems*, 29, 2016.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279. PMLR, 2019.

Margalit Glasgow, Colin Wei, Mary Wootters, and Tengyu Ma. Max-margin works while large margin fails: Generalization without uniform convergence. *arXiv preprint arXiv:2206.07892*, 2022.

Peter Grunwald, Thomas Steinke, and Lydia Zakynthinou. Pac-bayes, mac-bayes and conditional mutual information: Fast rate bounds that handle general vc classes. In *Conference on Learning Theory*, pages 2217–2247. PMLR, 2021.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

Hrayr Harutyunyan, Kyle Reing, Greg Ver Steeg, and Aram Galstyan. Improving generalization by controlling label-noise information in neural network weights. In *International Conference on Machine Learning*, pages 4071–4081. PMLR, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Wee Sun Lee, Peter L Bartlett, and Robert C Williamson. The importance of convexity in learning with squared loss. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 140–146, 1996.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

Yunwen Lei, Ting Hu, and Ke Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *The Journal of Machine Learning Research*, 22(1):1145–1185, 2021.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

Yi-An Ma, Teodor Vanislavov Marinov, and Tong Zhang. Dimension independent generalization of dp-sgd for overparameterized smooth convex optimization. *arXiv preprint arXiv:2206.01836*, 2022.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience, 1983.

Konstantinos E Nikolakakis, Farzin Haddadpour, Amin Karbasi, and Dionysios S Kalogerias. Beyond lipschitz: sharp generalization and excess risk bounds for full-batch gd. *arXiv preprint arXiv:2204.12446*, 2022.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. In *Conference on Learning Theory*, pages 250–296. PMLR, 2018a.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018b.

Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory*, pages 3380–3394. PMLR, 2022.

Ayush Sekhari, Karthik Sridharan, and Satyen Kale. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *Advances In Neural Information Processing Systems*, 34:27422–27433, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, volume 2, page 5, 2009.

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

Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. How does early stopping help generalization against label noise? *arXiv preprint arXiv:1911.08059*, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using gaussian kernels. *Annals of statistics*, 35(2):575–607, 2007.

Hossein Taheri and Christos Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *arXiv preprint arXiv:2302.09235*, 2023.

Jiaye Teng, Jianhao Ma, and Yang Yuan. Towards understanding generalization via decomposing excess risk dynamics. In *The Tenth International Conference on Learning Representations, ICLR 2022,* *Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

Kaiyue Wen, Jiaye Teng, and Jingzhao Zhang. Realistic deep learning may not fit benignly. *arXiv preprint arXiv:2206.00501*, 2022.

Jun Yang, Shengyang Sun, and Daniel M Roy. Fast-rate pac-bayes generalization bounds via shifted rademacher processes. *Advances in Neural Information Processing Systems*, 32, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pingali, Chao Chen, and Mayank Goswami. Stability of sgd: Tightness analysis and improved bounds. In *Uncertainty in Artificial Intelligence*, pages 2364–2373. PMLR, 2022.

Lijia Zhou, Frederic Koehler, Danica J Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *arXiv preprint arXiv:2112.04470*, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix: Proofs and Supplementary Materials

A Missing Proofs from Section 3.1

A.1 Proof of Theorem 3.1

The proof of GD is immediate from combining lower bound $\Omega\left(1/\eta T\right)$ in Lemma D.2, lower bound $\Omega\left(1/n\right)$ in Lemma D.1, and most importantly, lower bound $\Omega\left(\eta T/n^2\right)$ in Lemma A.1 to be stated below. In precise, Lemma A.1 is the core part of our result and gives a lower bound of $\Omega\left(\eta T/n\right)$ when $\eta T = O\left(n\right)$, and a lower bound of $\Omega\left(1/\eta T\right)$ when $\eta T = \Omega\left(n\right)$ for GD. The latter part is used in the proof of Theorem 3.2. We postpone its proof to Appendix A.3. The proof of the rest two lemmas can be found in Appendix D.1, D.2.

Lemma A.1. *For every $\eta > 0$, $T > 1$, if $\eta T = O\left(n\right)$, then there exists a convex, 1-smooth and realizable $f(W,Z)$: $\mathbb{R}^{(n+1)\times m} \times \mathcal{Z}^m \to \mathbb{R}$ for every $z \in \mathcal{Z}$, and a distribution D such that, with initialization $\|W_1 - W^*\| = O\left(1\right)$, the output W_{GD} for GD satisfies*

$$\mathbb{E}[F(W_{GD})] - F(W^*) = \Omega\left(\frac{\eta T}{n^2}\right).$$

In specific, m is an integer with $m = \Theta(e^n/\sqrt{n})$. Similarly, if $\eta T = \Omega\left(n\right)$, then it satisfies

$$\mathbb{E}[F(W_{GD})] - F(W^*) = \Omega\left(\frac{1}{\eta T}\right).$$

Similar to the proof of GD, the result on SGD is also obtained by combining lower bound constructions in the following lemma and Lemma D.1, D.2 in Appendix D.1, D.2. Lemma A.2 establishes a lower bound of $\Omega\left(\eta T/n\right)$ when $\eta T = O\left(n\right)$, and a lower bound of $\Omega\left(1/\eta T\right)$ when $\eta T = \Omega\left(n\right)$ for SGD. Its proof can be found in Appendix A.4.

Lemma A.2. *For every $\eta > 0$, $T > 1$, if $\eta T = O\left(n\right)$, then there exists a convex, 1-smooth and realizable $f(W,Z)$: $\mathbb{R}^{(n+1)\times m} \times \mathcal{Z}^m \to \mathbb{R}$ for every $Z \in \mathcal{Z}^m$, and a distribution D such that, with initialization $\|W_1 - W^*\| = O\left(1\right)$, the output W_{SGD} for SGD satisfies*

$$\mathbb{E}[F(W_{SGD})] - F(W^*) = \Omega\left(\frac{\eta T}{n^2}\right).$$

In specific, m is an integer with $m = \Theta(e^n/\sqrt{n})$. Similarly, if $\eta T = \Omega\left(n\right)$, then it satisfies

$$\mathbb{E}[F(W_{SGD})] - F(W^*) = \Omega\left(\frac{1}{\eta T}\right).$$

A.2 Proof of Theorem 3.2

Similar to the proof of Theorem 3.1, the proof of GD is obtained from combining the lower bounds in Lemma D.1, Lemma D.2, and Lemma A.1. In particular, Lemma A.1 gives a lower bound of $\Omega\left(\eta T/n\right)$ when $\eta T = \Omega\left(n\right)$.

Concurrently, the proof of SGD is obtained from combining the lower bounds in Lemma D.1, Lemma D.2, and Lemma A.2. In particular, Lemma A.2 gives a lower bound of $\Omega\left(1/\eta T\right)$ when $\eta T = \Omega\left(n\right)$.

A.3 Proof of Lemma A.1

This subsection contains the proof of Lemma A.1, along with another two supportive lemmas. We first present the proof of the major lemma.

Proof. We construct the following instance to obtain the lower bound, where $f : \mathbb{R}^{(n+1)\times m} \times \mathcal{Z}^m \to \mathbb{R}$ is

$$f(W,Z) = \sum_{j=1}^{m} g(w^{(i)}, z^{(i)}) \tag{7}$$

with a positive integer m, and g defined as

$$g(w, z = i) = \frac{\alpha}{2}x^2 + \frac{1}{2}\big(y(i)\big)^2 - \sqrt{\alpha}x \cdot y(i) = \frac{1}{2}\Big(\sqrt{\alpha}x - y(i)\Big)^2, \tag{8}$$

where $W = (w^{(1)}, \cdots, w^{(m)})$ is a large vector formed by concatenating by m vectors $\{w^{(j)} \mid w^{(j)} \in \mathbb{R}^{n+1}\}_{j=1}^m$, and $Z = (z^{(1)}, \cdots, z^{(m)})$ denotes a large sample concatenated by m copies of independent samples $\{z^{(j)} \mid \ z^{(j)} \in [n]\}$. We will omit the upscript of j when it does not lead to confusion. Each w is split into $w = (x, y)$ with $x \in \mathbb{R}$ and $y \in \mathbb{R}^n$ in the function g, and we define $y(i)$ as the i-th coordinate of y. We assume $z \sim \mathrm{Unif}([n])$ i.i.d., and set parameters $m = \Theta(e^n/\sqrt{n})$, $\alpha = C/(\eta T)$, where $C \le 1$ is a constant. Intuitively, f can be regarded as the summation over m copies of $g(w^{(j)}, z^{(j)})$. Such a construction f satisfies the conditions in the statement of this lemma (see Lemma A.3 below).

Lemma A.4 (also see below) shows that: with constant probability, there exists at least one copy of $\{z_i^{(j)}\}_{i\in[n]}$ (for clarification, $z_i^{(j)}$ is the j-th component in the i-th sample Z_i within the dataset $S = \{Z_1, \ldots, Z_n\}$) satisfying

$$z_i^{(j)} = i, \quad \text{for all} \quad i \in [n],$$

without the loss of generality, we consider the identity permutation $\boldsymbol{\pi}(i) = i$. We use the following initialization:

$$x_1^{(k)} = \begin{cases} 1, & k = j, \\ 0, & k \ne j; \end{cases} \qquad \text{and} \qquad y_1^{(k)} = 0, \qquad \forall k \in [m].$$

We have then $\|W_1 - W^*\| = O(1)$. This allows us to focus on the j-th component only and hence we suppress the upscripts. In this context, the stochastic loss function g on this copy is written as

$$g(w, z_i) = \frac{\alpha}{2}(x)^2 + \frac{1}{2}\|y\|^2 - \frac{x\sqrt{\alpha}}{n}y(i), \qquad \forall i \in [n]. \tag{9}$$

From the above construction, GD formulates the following update

$$w_{t+1} = w_t - \frac{\eta}{n}\sum_{i=1}^n \nabla_w g(w_t, z_i)$$

with initialization $x_1 = 1$, $y_1 = 0$. The stochastic gradient is computed as

$$\nabla_x g(w, z_i) = \alpha x - \sqrt{\alpha}y(i), \qquad \nabla_y g(w, z_i) = (y(i) - \sqrt{\alpha}x) \cdot \boldsymbol{e}_i. \tag{10}$$

Since all coordinates in y are equivalent in the construction, we suppress the index of i and write $y_t = y_t(i)$ for any $i \in [n]$, $t \in [T]$. Then it formulates

$$x_{t+1} = x_t - \eta\alpha x_t + \frac{\eta\sqrt{\alpha}}{n}\sum_{i=1}^n y_t(i) = (1 - \alpha\eta)x_t + \eta\sqrt{\alpha}y_t,$$
$$y_{t+1} = y_t - \frac{\eta}{n}y_t + \frac{\eta\sqrt{\alpha}}{n}x_t = \Big(1 - \frac{\eta}{n}\Big)\, y_t + \frac{\eta\sqrt{\alpha}}{n}x_t.$$

We next provide both upper and lower bounds for x_t and y_t. We give an upper bound for x_t and y_t by the following induction. If condition

$$x_t \le 1, \qquad y_t \le \sqrt{\alpha} \tag{11}$$

holds for t, then the above condition also holds for $t + 1$:

$$x_{t+1} \le (1 - \alpha\eta) + \eta\sqrt{\alpha} \cdot \sqrt{\alpha} = 1 - \frac{\eta}{\eta T} + \frac{\eta}{\eta T} \le 1,$$
$$y_{t+1} \le \Big(1 - \frac{\eta}{n}\Big)\, \sqrt{\alpha} + \eta\frac{\sqrt{\alpha}}{n} = \sqrt{\alpha}.$$

Then by induction we conclude that (11) is true. For any $t \in [T]$ with $T \geq 2$, the lower bound for x_t is much simpler to compute under our choice of parameter $\alpha = C/(\eta T)$:

$$x_{t+1} \geq (1-\alpha\eta)x_t \geq (1-\alpha\eta)^t x_1 = 4^{-Ct/T} \geq 4^{-C}.$$

Hence $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t = \Theta(1)$. This then allows us to lower bound y at iteration $t \in [T]$:

$$\begin{aligned} y_t &\geq \left(1 - \frac{\eta}{n}\right) y_{t-1} + \frac{\eta\sqrt{\alpha}}{4^C n} \\ &\geq \frac{\eta\sqrt{\alpha}}{4^C n} \cdot \left(1 + (1-\eta/n) + \cdots (1-\eta/n)^{t-1}\right) \\ &\geq \frac{\eta\sqrt{\alpha}}{4^C n} \cdot \frac{1-(1-\eta/n)^t}{1-(1-\eta/n)}. \end{aligned}$$

Now, we discuss two cases: $\eta T = O(n)$ and $\eta T = \Omega(n)$.

Case $\eta T = O(n)$. We decompose $t = n \cdot \frac{t}{n}$ and obtain

$$\begin{aligned} y_t &\geq \frac{\eta\sqrt{\alpha}}{4^C n} \cdot \frac{1-(1-\eta/n)^t}{1-(1-\eta/n)} = \frac{\eta\sqrt{\alpha}}{4^C n} \cdot \frac{1-(1-\eta/n)^{\frac{t}{n}\cdot n}}{1-(1-\eta/n)} \\ &\overset{\text{(A)}}{\geq} \frac{\eta\sqrt{\alpha}}{4^C}\left(\frac{t}{n} - \frac{\eta t^2}{2n^2}\right) \overset{\text{(B)}}{=} \frac{\eta t\sqrt{\alpha}}{2\cdot 4^C n} = \sqrt{\frac{\eta}{CT}} \cdot \frac{t}{2\cdot 4^C n} \end{aligned}$$

where (A) is due to Taylor expansion, (B) is due to the condition $\eta t \leq \eta T = O(n)$ and $\alpha = C/(\eta T)$. We then calculate the average output

$$\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{\eta}{CT}} \cdot \frac{t}{2\cdot 4^C n} \geq \frac{1}{4\cdot 4^C n} \cdot \sqrt{\frac{\eta T}{C}}.$$

We return to the original $f(w,z)$ by inserting the above analysis on the j-th component:

$$\mathbb{E}[F(W_{\text{GD}})] \geq \mathbb{E}[G(w_{\text{GD}}^{(j)})] \geq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{\text{GD}}^{(j)}}{\sqrt{\eta T}} - y_{\text{GD}}^{(j)}(i)\right)^2 \geq \Omega\left(\max\left\{\frac{1}{\eta T}, \frac{\eta T}{n^2}\right\}\right) = \Omega\left(\frac{\eta T}{n^2}\right)$$

where the last inequality is due to the fact $4^{-C} \leq x_{\text{GD}} \leq 1$. We can always choose a proper C such that the difference is non-vanishing.

Case $\eta T = \Omega(n)$. We can directly lower bound y_t as

$$y_t \geq \frac{\eta\sqrt{\alpha}}{4^C n} \cdot \frac{n}{2\eta} = \frac{1}{2\cdot 4^C\sqrt{C\eta T}} = \Omega\left(\frac{1}{\sqrt{\eta T}}\right)$$

since $(1-\eta/n)^t \leq 1/2$ when $\eta T = \Omega(n)$. We then calculate the average output

$$\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t = \frac{1}{T}\sum_{t=1}^{T}\Omega\left(\frac{1}{\sqrt{\eta T}}\right) \geq \Omega\left(\frac{1}{\sqrt{\eta T}}\right).$$

Similarly, we return to the original $f(w,z)$ by inserting the above analysis on the j-th component and obtain a non-vanishing lower bound by choosing proper C:

$$\mathbb{E}[F(W_{\text{GD}})] \geq \mathbb{E}[G(w_{\text{GD}}^{(j)})] \geq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{\text{GD}}^{(j)}}{\sqrt{\eta T}} - y_{\text{GD}}^{(j)}(i)\right)^2 \geq \Omega\left(\frac{1}{\eta T}\right).$$

This completes our proof. □

We proceed to prove the supporting lemmas.

Lemma A.3. *Suppose $\alpha = \Theta(1/\eta T)$, then f is 1-smooth, convex and realizable over D.*

Proof. When condition $\eta T \geq 1$ holds, it is easy to check that $g(w, z)$ is 1-smooth and convex for any $w \in \mathbb{R}^{n+1}$ and $z \in [n]$. The population risk G is

$$G(w) = \mathbb{E}_{z \sim \mathrm{Unif}([n])}[g(w, z)] = \frac{\alpha}{2}x^2 + \frac{1}{2n}\|y\|^2 - \frac{\sqrt{\alpha}}{n}x \cdot \mathbf{1}^\top y = \frac{1}{2n}\left\|y - \sqrt{\alpha}x \cdot \mathbf{1}\right\|^2,$$

which attains minimum at $(x^*, y^*) = (0, 0)$. So $g(w, z)$ satisfies the realizable condition. It is easy to conclude f is also 1-smooth, convex and realizable. □

Lemma A.4. *Consider dataset $S = \{Z_1, \ldots, Z_n\}$ defined in Lemma A.1. Suppose m is a positive integer satisfying $m = \Theta(e^n/\sqrt{n})$, then there exists at least one component $z^{(j)}, j \in [m]$ such that*

$$z_i^{(j)} = \boldsymbol{\pi}(i), \quad \textit{for all} \quad i \in [n]$$

where $\boldsymbol{\pi} : [n] \to [n]$ is any permutation on $[n]$.

Proof. We define the following probability event: given dataset $S = \{Z_1, \ldots, Z_n\}$, we focus on the j-th component $\{z_1^{(j)}, \ldots, z_n^{(j)}\}$ and define event $\mathcal{E}_j$ as

$$\mathcal{E}_j = \left\{ z_i^{(j)} = \boldsymbol{\pi}(i) \text{ for any } i \in [n] \right\}$$

where $\boldsymbol{\pi} : [n] \to [n]$ is any fixed permutation on $[n]$. Intuitively, when $\mathcal{E}_j$ happens, each coordinates of $y^{(j)}$ is selected only for once in dataset S. For any fixed $j \in [m]$, the probability of $\mathcal{E}_j$ happens is calculated from the without-replacement sampling:

$$p := \Pr[\text{Event } \mathcal{E}_j \text{ happens}] = 1 \cdot \frac{n-1}{n} \cdot \cdots \cdot \frac{1}{n} = \frac{n!}{n^n} = \Theta(\sqrt{n} \cdot e^{-n})$$

where the last step is from Stirling approximation $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$ for any $n \geq 1$. ensures event $\mathcal{E}_j$ to happen.

We now prove that with $\Omega(1)$ probability, there exists at least a $j \in [m]$ such that $\mathcal{E}_j$ happens using the second moment method. Denote R to be the random variable counting the number of $\{\mathcal{E}_j\}_{j \in [m]}$ happens. Using second moment method, we upper bound the following probability:

$$\Pr[R > 0] \geq \frac{(\mathbb{E}[R])^2}{\mathbb{E}[R^2]} = \frac{m^2 p^2}{mp(1-p)} \geq \frac{mp}{2} = \frac{1}{2}. \tag{12}$$

So with probability $\Omega\left(\frac{1}{2}\right)$, we have at least one copy fulfilling the statement. □

A.4 Proof of Lemma A.2

Proof. We utilize the similar strategy employed in Lemma A.1 and consider the same $f(W, Z)$ defined in Eq. (7), Eq.(8). Lemma A.4 (see below) shows that: with constant probability, there exists at least one copy of $\{z_i^{(j)}\}_{i \in [n]}$ satisfying (without the loss of generality, we consider the identical permutation $\boldsymbol{\pi}(i) = i$)

$$z_i^{(j)} = i, \quad \text{for all} \quad i \in [n].$$

We use the following initialization:

$$x_1^{(k)} = \begin{cases} 1, & k = j, \\ 0, & k \neq j; \end{cases} \qquad \text{and} \qquad y_1^{(k)} = 0, \qquad \forall k \in [m].$$

We have then $\|W_1 - W^*\| = O(1)$. This allows us to focus on the j-th component only and hence we suppress the upscripts. In this context, the stochastic loss function g on this copy is written as

$$g(w, z_i) = \frac{\alpha}{2}(x)^2 + \frac{1}{2}\|y\|^2 - \frac{x\sqrt{\alpha}}{n}y(i), \qquad \forall i \in [n]. \tag{13}$$

SGD formulates the update:

$$w_{t+1} = w_{t+1} - \eta g(w_t, z_{i_t})$$

where $z_{i_t} \sim \mathrm{Unif}([n])$. The initialization is $x_1 = 1$, $y_1 = 0$. Based on the current value of w_t, under expectation, we have

$$\mathbb{E}[w_{t+1}] = w_t - \frac{\eta}{n}\sum_{i=1}^{n} \nabla_w g(w_t, z_i).$$

We write the update of $\mathbb{E}[x_t]$ and $\mathbb{E}[y_t]$ by plugging stochastic gradients: it easy to see all coordinates in $\mathbb{E}[y_t]$ are equivalent, we suppress the index of i and write $y_t = y_t(i)$ for any $i \in [n]$, $t \in [T]$. Then it formulates

$$\mathbb{E}[x_{t+1}] = x_t - \eta\alpha x_t + \frac{\eta\sqrt{\alpha}}{n}\sum_{i=1}^{n} y_t(i) = (1-\alpha\eta)x_t + \eta\sqrt{\alpha}y_t,$$
$$\mathbb{E}[y_{t+1}] = y_t - \frac{\eta}{n}y_t + \frac{\eta\sqrt{\alpha}}{n}x_t = \left(1 - \frac{\eta}{n}\right)y_t + \frac{\eta\sqrt{\alpha}}{n}x_t.$$

We give an upper bound for $\mathbb{E}[x_t]$ and $\mathbb{E}[y_t]$ by the following induction. If condition

$$\mathbb{E}[x_t] \le 1, \qquad \mathbb{E}[y_t] \le \sqrt{\alpha} \tag{14}$$

holds for t, then the above condition also holds for $t+1$:

$$\mathbb{E}[x_{t+1}] \le (1-\alpha\eta) + \eta\sqrt{\alpha}\cdot\sqrt{\alpha} = 1 - \frac{\eta}{\eta T} + \frac{\eta}{\eta T} \le 1,$$
$$\mathbb{E}[y_{t+1}] \le \left(1 - \frac{\eta}{n}\right)\sqrt{\alpha} + \eta\frac{\sqrt{\alpha}}{n} = \sqrt{\alpha}.$$

Then by induction we conclude that (14) is true. For any $t \in [T]$ and $T \ge 2$, the lower bound for x_t is much simpler to compute under our choice of parameter $\alpha = C/(\eta T)$:

$$\mathbb{E}[x_{t+1}] \ge (1-\alpha\eta)x_t \ge (1-\alpha\eta)^t x_1 = 4^{-t/T} \ge 4^{-C}.$$

Hence $\mathbb{E}[\bar{x}_T] = \frac{x_1}{T} + \frac{1}{T}\sum_{t=2}^{T}\mathbb{E}[x_t|w_{t-1}] = \Theta(1)$. This then allows us to lower bound y at iteration $t \in [T]$:

$$\begin{aligned}\mathbb{E}[y_t] &\ge \left(1 - \frac{\eta}{n}\right)y_{t-1} + \frac{\eta\sqrt{\alpha}}{4^C n} \\ &\ge \frac{\eta\sqrt{\alpha}}{4^C n}\cdot\left(1 + (1-\eta/n) + \cdots (1-\eta/n)^{t-1}\right) \\ &\ge \frac{\eta\sqrt{\alpha}}{4^C n}\cdot\frac{1-(1-\eta/n)^t}{1-(1-\eta/n)}.\end{aligned}$$

Now, we discuss two cases: $\eta T = O(n)$ and $\eta T = \Omega(n)$.

Case $\eta T = O(n)$. We decompose $t = n \cdot \frac{t}{n}$ and obtain

$$\begin{aligned}\mathbb{E}[y_t] &\ge \frac{\eta\sqrt{\alpha}}{4^C n}\cdot\frac{1-(1-\eta/n)^t}{1-(1-\eta/n)} = \frac{\eta\sqrt{\alpha}}{4^C n}\cdot\frac{1-(1-\eta/n)^{\frac{t}{n}\cdot n}}{1-(1-\eta/n)} \\ &\overset{\text{(A)}}{\ge} \frac{\eta\sqrt{\alpha}}{4^C}\left(\frac{t}{n} - \frac{\eta t^2}{2n^2}\right) \overset{\text{(B)}}{=} \frac{\eta t\sqrt{\alpha}}{2\cdot 4^C n} = \sqrt{\frac{\eta}{CT}}\cdot\frac{t}{2\cdot 4^C n}\end{aligned}$$

where (A) is due to Taylor expansion, (B) is due to the condition $\eta t \le \eta T = O(n)$ and $\alpha = C/(\eta T)$. We then calculate the average output

$$\mathbb{E}[\bar{y}_T] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[y_t] = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{\eta}{CT}}\cdot\frac{t}{2\cdot 4^C n} \ge \frac{1}{4\cdot 4^C n}\cdot\sqrt{\frac{\eta T}{C}}.$$

We return to the original $f(w,z)$ by inserting the above analysis on the j-th component:

$$\begin{aligned}\mathbb{E}[F(W_{\mathrm{SGD}})] \ge \mathbb{E}[G(w_{\mathrm{SGD}}^{(j)})] \ge G(\mathbb{E}[w_{\mathrm{SGD}}^{(j)}]) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{\mathrm{SGD}}^{(j)}}{\sqrt{\eta T}} - y_{\mathrm{SGD}}^{(j)}(i)\right)^2 \\ &\ge \Omega\left(\max\left\{\frac{1}{\eta T,}\ \frac{\eta T}{n^2}\right\}\right) = \Omega\left(\frac{\eta T}{n^2}\right)\end{aligned}$$

where the second inequality comes from Jensen's inequality and the last inequality is due to the fact $4^{-C} \le x_{\mathrm{GD}} \le 1$. We can always choose a proper C such that the difference is non-vanishing.

Case $\eta T = \Omega(n)$. We can directly lower bound $\mathbb{E}[y_t]$ as

$$\mathbb{E}[y_t] \ge \frac{\eta\sqrt{\alpha}}{4^C n}\cdot\frac{n}{2\eta} = \frac{1}{2\cdot 4^C\sqrt{C\eta T}} = \Omega\left(\frac{1}{\sqrt{\eta T}}\right)$$

since $(1-\eta/n)^t \le 1/2$ when $\eta T = \Omega(n)$. We then calculate the average output

$$\mathbb{E}[\bar{y}_T] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[y_t] = \sum_{t=1}^{T}\Omega\left(\frac{1}{\sqrt{\eta T}}\right) \ge \Omega\left(\frac{1}{\sqrt{\eta T}}\right).$$

Similarly, we return to the original $f(w,z)$ by inserting the above analysis on the j-th component and obtain a non-vanishing lower bound by choosing proper C:

$$\mathbb{E}[F(W_{\mathrm{SGD}})] \ge \mathbb{E}[G(w_{\mathrm{SGD}}^{(j)})] \ge G(\mathbb{E}[w_{\mathrm{SGD}}^{(j)}]) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{\mathrm{SGD}}^{(j)}}{\sqrt{\eta T}} - y_{\mathrm{SGD}}^{(j)}(i)\right)^2 \ge \Omega\left(\frac{1}{\eta T}\right).$$

This completes our proof. □

B Missing Proofs from Section 3.2

B.1 Proof of Theorem 3.3

The theorem provides an excess risk lower bound $\Omega(1/\eta T + \eta T/n)$ for GD under the *non-realizable* smooth SCO scenario. The result is obtained by combining a $\Omega(1/\eta T)$ bound in Lemma D.2 and a $\Omega(\eta T/n)$ bound in Lemma B.1 stated below. The first bound reflects an optimization error and is postponed to Appendix D.2. In the rest part, we present the proof of the latter lemma.

Lemma B.1. *For any $\eta > 0$, $T > 1$, there exists a convex, 1-smooth $f(w,z) : \mathbb{R} \to \mathbb{R}$ for every $z \in \mathcal{Z}$, and a distribution D such that, with probability $\Theta(1)$, the output w_{GD} for GD satisfies*

$$F(w_{GD}) - F(w^*) = \Omega\left(\frac{\eta T}{n}\right).$$

Proof. We define loss function $f : \mathbb{R}\times\mathcal{Z} \to \mathbb{R}$ as

$$f(w,z) = \frac{w^2}{2\eta T} + zw$$

where $z \sim \mathrm{Unif}(\{\pm 1\})$. It is obvious that $f(w,z)$ is 1-smooth and convex since $\eta T \ge 1$. The population risk is computed as

$$F(w) = \mathbb{E}_{z\sim\mathrm{Unif}(\{\pm1\})}[f(w,z)] = \frac{w^2}{2\eta T}.$$

The minimizer is then $w^* = 0$. GD formulates the following recurrence on dataset S with initialization $w_1 = 0$:

$$w_{t+1} = w_t - \frac{\eta}{n}\sum_{i=1}^{n}\left(\frac{w_t}{\eta T} + z_i\right) = \left(1 - \frac{1}{T}\right)w_t - \frac{\eta}{n}\sum_{i=1}^{n} z_i,$$

where each $z_i \sim \text{Unif}(\{\pm 1\})$ for $i \in [n]$. We want to use an anti-concentration result to lower bound the recurrence: from Lemma 7 in Sekhari et al. (2021), with probability $\Omega(1)$, it holds that

$$\sum_{i=1}^{n} z_i \leq -\frac{\sqrt{n}}{2}.$$

We get lower bound

$$w_{t+1} \geq \left(1 - \frac{1}{T}\right)w_t + \frac{\eta}{2\sqrt{n}}.$$

Then we have for any $t \in [T]$

$$\begin{aligned} w_t &\geq \frac{\eta}{2\sqrt{n}}\left(1 + \left(1 - \frac{1}{T}\right) + \cdots + \left(1 - \frac{1}{T}\right)^{t-1}\right) \\ &\geq \frac{\eta t}{8\sqrt{n}} \end{aligned}$$

where the second inequality is due to the fact

$$1 > 1 - \frac{1}{T} > \cdots \left(1 - \frac{1}{T}\right)^{t} > \cdots > \left(1 - \frac{1}{T}\right)^{T} \geq \frac{1}{4}$$

for any $t \in [T]$ and $T \geq 2$. Then the average is lower bounded as

$$\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t \geq \sum_{t=1}^{T}\frac{\eta t}{8\sqrt{n}} = \frac{\eta(T-1)}{16\sqrt{n}}.$$

As a result, we have

$$F(w_{\text{GD}}) - F(w^*) = \frac{w_{\text{GD}}^2}{2\eta T} = \Omega\left(\frac{\eta T}{n}\right),$$

which is the desired result. □

B.2 Proof of Theorem 3.4

Similar to the proof of Theorem 3.3, we prove the excess risk lower bound for SGD by combining Lemma D.2 and the following lemma.

Lemma B.2. *For any $\eta > 0$, $T > 1$, there exists a convex, 1-smooth $f(w, z) : \mathbb{R} \to \mathbb{R}$ for every $z \in \mathcal{Z}$, and a distribution D such that, with probability $\Theta(1)$, the output w_{SGD} for SGD satisfies*

$$\mathbb{E}[F(w_{SGD})] - F(w^*) = \Omega\left(\frac{\eta T}{n}\right).$$

Proof. We use the same construction in Lemma B.1. Consider dataset $S = \{z_1, \ldots, z_n\}$ where $z_i \sim \text{Bern}(\{\pm 1\})$. Given x_t, SGD formulates the following recurrence on dataset S with initialization $w_1 = 0$:

$$\mathbb{E}[w_{t+1}] = w_t - \frac{\eta}{n}\sum_{i=1}^{n}\left(\frac{w_t}{\eta T} + z_i\right) = \left(1 - \frac{1}{T}\right)w_t - \frac{\eta}{n}\sum_{i=1}^{n} z_i,$$

where $z_i \sim \mathrm{Unif}(\{\pm 1\})$. From Lemma 7 in Sekhari et al. (2021), with probability $\Omega(1)$, it holds that

$$\sum_{i=1}^{n} z_i \leq -\frac{\sqrt{n}}{2}.$$

Then we get lower bound

$$\mathbb{E}[w_{t+1}] \geq \left(1 - \frac{1}{T}\right) w_t + \frac{\eta}{2\sqrt{n}}.$$

Similar to the proof of Lemma B.1, we have for any $t \in [T]$

$$\begin{aligned}\mathbb{E}[w_t] &\geq \frac{\eta}{2\sqrt{n}} \left(1 + \left(1 - \frac{1}{T}\right) + \cdots + \left(1 - \frac{1}{T}\right)^{t-1}\right) \\ &\geq \frac{\eta t}{8\sqrt{n}}.\end{aligned}$$

Then the average is lower bounded as

$$\mathbb{E}[\bar{w}_T] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[w_t] \geq \sum_{t=1}^{T} \frac{\eta t}{8\sqrt{n}} = \frac{\eta(T-1)}{16\sqrt{n}}.$$

As a result, we have

$$\mathbb{E}[F(w_{\mathrm{SGD}})] - F(w^*) \geq F(\mathbb{E}[w_{\mathrm{SGD}}]) - F(w^*) = \frac{(\mathbb{E}[w_{\mathrm{SGD}}])^2}{2\eta T} = \Omega\left(\frac{\eta T}{n}\right),$$

by Jensen's inequality. □

C Missing Proofs in Section 4

Here we provide proofs of Lemma 4.1 and Lemma 4.2, in dimension one, SGD/GD are also able to achieve $O(1/n)$ sample complexity under the regime $T = \Omega(n)$.

Lemma C.1 (Restated Lemma 4.1)**.** *In dimension one, if $f(w, z)$ is convex, 1-smooth and realizable with $z \sim D$, then for every $\eta = \Theta(1)$, there exists $T_0 = \Theta(n)$ such that for $T \geq T_0$, the output w_{SGD} of SGD satisfies*

$$\mathbb{E}[F(w_{SGD})] - F(w^*) = O\left(\frac{1}{n}\right).$$

Proof. From Theorem 4 in Lei and Ying (2020), it holds that for realizable cases (we rescale it to $f(w^*, z) = 0$ for each z) with step-size $\eta = \Theta(1)$, it holds that

$$\mathbb{E}[F(w_{\mathrm{SGD}})] = O\left(\frac{1}{T_0} + \frac{1 + T_0/n}{n}\right). \tag{15}$$

Therefore, for $T_0 = \Theta(n)$, it holds that

$$\mathbb{E}[F(w_{\mathrm{SGD}})] = O(1/n).$$

For SGD, the iteration formulates the iterate

$$w_{t+1} = w_t - \eta \nabla f(w_t, z_{i_t}),$$

where z_{i_t} is uniformly chosen from S. Under the realizable and convex assumption, for any $z_{i_t} \in \mathcal{Z}$, the iteration becomes

$$w_{t+1} - w^* = (1 - \eta \nabla^2 f(\xi, z_{i_t}))(w_t - w^*),$$

using mean value theorem, where ξ is a point between w_t and w^*. This indicates that the distance $w_t - w^*$ shrinks in each step for any $z_{i_t} \in \mathcal{Z}$. Due to the convexity of F, it holds that $F(w_{t+1}) \leq F(w_t)$. In summary, for any $T \geq T_0$, it holds that

$$\mathbb{E}[F(w_{\mathrm{SGD}})] - F(w^*) = O(1/n).$$

□

Lemma C.2 (Restated Lemma 4.2)**.** *In dimension one, if $f(w,z)$ is convex, 1-smooth and realizable with $z \sim D$, then for every $\eta = \Theta(1)$, there exists $T_0 = \Theta(n)$ such that for $T \geq T_0$, the output w_{GD} of GD satisfies*

$$\mathbb{E}[F(w_{GD})] - F(w^*) = O\left(\frac{1}{n}\right).$$

Proof. From Theorem 10 in Nikolakakis et al. (2022), it holds that for realizable cases (we rescale it to $f(w^*, z) = 0$ for each z) with step-size $\eta = \Theta(1)$, it holds that

$$\mathbb{E}[F(w_{\mathrm{GD}})] = O\left(\frac{1}{T_0} + \frac{1 + T_0/n}{n}\right). \tag{16}$$

Therefore, for $T_0 = \Theta(n)$, it holds that

$$\mathbb{E}[F(w_{\mathrm{GD}})] = O\left(1/n\right).$$

For SGD, the iteration formulates the iterate

$$w_{t+1} = w_t - \eta \nabla F_S(w_t).$$

Under the realizable and convex assumption, the iteration becomes

$$w_{t+1} - w^* = (1 - \eta \nabla^2 F_S(\xi))(w_t - w^*),$$

using mean value theorem, where ξ is a point between w_t and w^*. This indicates that the distance $w_t - w^*$ shrinks in each step. Due to the convexity of F, it holds that $F(w_{t+1}) \leq F(w_t)$. In summary, for any $T \geq T_0$, it holds that

$$\mathbb{E}[F(w_{\mathrm{GD}})] - F(w^*) = O\left(1/n\right).$$

which is the desired result. □

D Minor Proofs

D.1 Lower Bound of Term $1/n$

Lemma D.1. *For every $\eta > 0$, $T > 1$, there exists a convex, 1-smooth and realizable $f(w,z) : \mathbb{R}^{2n} \to \mathbb{R}$ for every $z \in \mathcal{Z}$, and a distribution D such that, it holds for the output of any gradient-based algorithm $\mathcal{A}[S]$*

$$\mathbb{E}[F(\mathcal{A}[S])] - F(w^*) = \Omega\left(1/n\right).$$

Proof. We consider the following instance

$$f(w, z = i) = \frac{1}{2} w(i)^2, \qquad z \sim \mathrm{Uniform}([2n]),$$

where $w(i)$ denotes the i-th coordinate of w. Then the population risk is

$$F(w) = \mathbb{E}_{z \sim \mathrm{Uniform}([2n])}[f(w,z)] = \frac{1}{4n}\|w\|^2,$$

which achieves minimum at $w^* = 0$. It is easy to check that $f(w,z)$ is 1-smooth, convex and realizable. Now consider any dataset S of n samples. Since $z \sim \mathrm{Uniform}([2n])$, with probability $\Omega(1)$, $\Theta(n)$ coordinates are not observed. For any gradient-based algorithm with initialization $w_0 = \frac{1}{\sqrt{2n}} \cdot \mathbf{1}_d$, the unobserved $\Theta(n)$ coordinates will remain unchanged for any step-size η and T. Then we have the following lower bound:

$$\mathbb{E}[F(\mathcal{A}[S])] - F(w^*) = \Omega\left(\frac{1}{4n} \cdot \frac{n}{2n}\right) = \Omega\left(\frac{1}{n}\right),$$

which is the desired result. □

D.2 Lower Bound of Term $1/\eta T$

Lemma D.2. *For every $\eta > 0$, $T > 1$, there exists a convex, 1-smooth and realizable $f(w, z) : \mathbb{R}^2 \to \mathbb{R}$ for every $z \in \mathcal{Z}$, and a distribution D such that, the output w_{GD} for GD satisfies*

$$\mathbb{E}[F(w_{GD})] - F(w^*) = \Omega\left(\frac{1}{\eta T}\right).$$

The same result also holds for SGD.

Proof. As usual we suppose $\eta T \geq 1$. We define the deterministic convex and 1-smooth function as

$$f(w) = \frac{1}{2}w^2(1) + \frac{\lambda}{2}w^2(2) \tag{17}$$

with $0 < \lambda < 1$, $w(1)$ and $w(2)$ are the value of first and second coordinate of w. Then GD formulates the iteration

$$w_{t+1} = w_t - \eta\nabla f(w_t),$$

with initialization $w_1 = (1, 1)$. This is then precisely:

$$w_{t+1}(1) = (1-\eta)w_{t+1}(1), \qquad w_{t+1}(2) = (1-\lambda\eta)w_{t+1}(2).$$

With $\lambda = \frac{1}{\eta T}$, we can upper bound for $t \in [T-1]$:

$$w_{t+1}(1) \geq \frac{1}{4}e^{-\eta t} \cdot x_1(1), \qquad w_{t+1}(2) \geq \frac{1}{4}e^{-\lambda\eta t} \cdot x_1(2) = \frac{e^{-t/T}}{4}.$$

The averaged output is lower bounded as

$$\bar{w}_T(2) = \sum_{t=1}^{T} \bar{w}_t(2) \geq \sum_{t=1}^{T} \frac{e^{-(t-1)/T}}{4T} \geq \frac{1}{4e} > \frac{1}{12}.$$

where the second inequality is due to the fact

$$1 > e^{-1/T} > \cdots > e^{-t/T} > \cdots > e^{-T/T} = e^{-1}$$

for any $t \in [T]$. Therefore, the suboptimality is

$$f(w_{\mathrm{GD}}) - f(w^*) \geq \frac{\lambda}{2}|w_{\mathrm{GD}}(2)|^2 \geq \frac{1}{288\eta T}. \tag{18}$$

Then the following result holds:

$$\mathbb{E}[F(w_{\mathrm{GD}})] - F(w^*) = f(w_T) - f(w^*) \geq \Omega\left(\frac{1}{\eta T}\right)$$

because $f(w)$ is a deterministic function. Since the instance is deterministic, then the suboptimality lower bound $\Omega\left(\frac{1}{\eta T}\right)$ also holds for SGD. □

D.3 GD Upper Bound for Realizable Smooth SCO

Here we derive the upper bound for GD under realizable smooth SCO, as in Table 1. The derivation is based on Theorem 10 in Nikolakakis et al. (2022). In the realizable cases, it holds that (see Nikolakakis et al. (2022) for the notations):

$$\epsilon_{\mathrm{opt}} = \frac{\mathbb{E}\|w_1 - w_S^*\|^2}{\eta T}$$
$$\epsilon_{\mathrm{path}} = \beta(\mathbb{E}\|w_1 - w_S^*\|^2 + \epsilon_{\boldsymbol{c}}\eta T)$$
$$\epsilon_{\boldsymbol{c}} = 0$$

Plug them in Theorem 10, it holds that for some constant c,

$$|\epsilon_{\mathrm{gen}}| \le c\frac{\beta\mathbb{E}\|w_1 - w_S^*\|^2}{n} + \frac{\beta^2\eta T\mathbb{E}\|w_1 - w_S^*\|^2}{n^2}.$$

Combined with the optimization upper bound $O\left(1/\eta T\right)$, we obtain the upper bound

$$\mathbb{E}[F(w_{\mathrm{GD}})] - F(w^*) = O\left(\frac{1}{\eta T} + \frac{1}{n} + \frac{\eta T}{n^2}\right).$$