Tight Risk Bounds for Gradient Descent on Separable Data

Matan Schliserman[1][*]   **Tomer Koren**[1][2]

[1]Blavatnik School of Computer Science, Tel Aviv University
[2]Google Research, Tel Aviv

Abstract

We study the generalization properties of unregularized gradient methods applied to separable linear classification—a setting that has received considerable attention since the pioneering work of Soudry et al. [14]. We establish tight upper and lower (population) risk bounds for gradient descent in this setting, for any smooth loss function, expressed in terms of its tail decay rate. Our bounds take the form $\Theta(r_{\ell,T}^2/\gamma^2 T + r_{\ell,T}^2/\gamma^2 n)$, where T is the number of gradient steps, n is size of the training set, γ is the data margin, and $r_{\ell,T}$ is a complexity term that depends on the tail decay rate of the loss function (and on T). Our upper bound greatly improves the existing risk bounds due to Shamir [13] and Schliserman and Koren [12], that either applied to specific loss functions or imposed extraneous technical assumptions, and applies to virtually any convex and smooth loss function. Our risk lower bound is the first in this context and establish the tightness of our general upper bound for any given tail decay rate and in all parameter regimes. The proof technique used to show these results is also markedly simpler compared to previous work, and is straightforward to extend to other gradient methods; we illustrate this by providing analogous results for Stochastic Gradient Descent.

1   Introduction

Recently, there has been a marked increase in interest regarding the generalization capabilities of unregularized gradient-based learning methods. One specific area of attention in this context has been the setting of linear classification with separable data, where a pioneering work by Soudry et al. [14] showed that, when using plain gradient descent to minimize the empirical risk on a linearly separable training set with an exponentially-tailed classification loss (such as the logistic loss), the trained predictor will asymptotically converge in direction to the max-margin solution. As a result, standard margin-based generalization bounds for linear predictors suggest that, provided the number of gradient steps (T) is sufficiently large, the produced solution will not overfit, despite the lack of explicit regularization and the fact that its magnitude (i.e., Euclidean norm) increases indefinitely with T. This result has since been extended to incorporate other optimization algorithms and loss functions [3, 4, 8, 9, 5].

Despite the high interest in this problem, the tight finite-time (population) risk performance of unregularized gradient methods, and even just of gradient descent, have not yet been fully understood. The convergence to a high margin solution exhibited by Soudry et al. [14] (for the logistic loss) occurs at a slow logarithmic rate, thus, the risk bounds for the trained predictors only become effective when T is at least exponentially large in comparison to the size of the training set n and the margin γ. In a more recent work, Shamir [13] established risk bounds of the form $\widetilde{O}(1/\gamma^2 T + 1/\gamma^2 n)$ for several gradient methods in this setting (still with the logistic loss), that apply for smaller and more realistic values of T.

---

[*]Corresponding author; `schliserman@mail.tau.ac.il`.

Table 1: Comparison between the bounds established in this work and bounds established in previous work for for gradient descent on γ-separable data; here, T is the number of gradient steps and n is the size of the training set. Logarithmic factors other than $\log T$ factors are suppressed from the bounds. Examples of how our general bounds are instantiated for some concrete loss functions are presented in Table 2 below.

Paper	Class of losses	Type of bound	Upper bound	Lower bound
Shamir [13]	Logistic loss	0-1 loss (population)	$\frac{\log^2 T}{\gamma^2 T} + \frac{1}{\gamma^2 n}$	$\frac{1}{\gamma^2 T}$ (a)
Schliserman and Koren [12]	General convex; smooth; Lipschitz; (c,δ)-self-bounded;[(b)] Decay rate of ϕ	Risk (population)	$\frac{r_{\phi,T}^2}{\gamma^2 T} + \frac{T^\delta r_{\phi,T}^{2-2\delta}}{\gamma^2 n}$	–
	General convex; smooth; (c,δ)-self-bounded;[(b)] Decay rate of ϕ	Risk (population)	$\frac{r_{\phi,T}^2}{\gamma^2 T} + \frac{T^{2\delta} r_{\phi,T}^{4-4\delta}}{\gamma^4 n^{1+2\delta}}$	–
Telgarsky [16]	Convex linear models; quadratically-bounded;[(c)] Decay rate of ϕ	Risk (population)	$\frac{r_{\phi,T}}{\gamma\sqrt{T}}$ (d)	–
This paper	Convex linear models; smooth; Decay rate of ϕ	Risk (population)	$\frac{r_{\phi,T}^2}{\gamma^2 T} + \frac{r_{\phi,T}^2}{\gamma^2 n}$	$\frac{r_{\phi,T}^2}{\gamma^2 T} + \frac{r_{\phi,T}^2}{\gamma^2 n}$

[a] The lower bound is shown for the empirical (logistic) risk.
[b] f is (c,δ)-self-bounded if for every x, it holds that $\|\nabla f(x)\| \leq c f(x)^{1-\delta}$; for $\delta < 1/2$, this property is stronger than smoothness of f.
[c] ℓ is (c_1, c_2)-quadratically-bounded if, for every x and z, $|\ell'(x,z)| \leq c_1 + c_2(\|x\| + \|z\|)$.
[d] The upper bound is valid when $T \leq n$.

Later Schliserman and Koren [12] gave a more general analysis that extends the bounds of Shamir [13] to a wide range of smooth loss functions satisfying a certain "self-boundedness" condition, and Telgarsky [16] provided a test-loss analysis of stochastic mirror descent on "quadratically-bounded" losses. However, all of these works either assume a specific loss (e.g., the logistic loss), impose various conditions on the loss function (beyond smoothness), or do not establish the tightness of their bounds in all regimes of parameters $(T, n$ and $\gamma)$, and especially in the regime where the number of gradient steps T is larger than the sample size n (see Table 1 for details).

1.1 Our contributions

In this work, we close these gaps by showing nearly matching upper and lower risk bounds for gradient descent, which hold essentially for *any* smooth loss function, and for any number of steps T and sample size n. Compared to recent prior work in this context [13, 16, 12], our results do not require any additional assumptions on the loss function besides its smoothness, and they strictly improve upon the existing bounds by their dependence on T. Further, to the best of our knowledge, risk lower bounds were not previously explored in this context, and the lower bounds we give establish, for the first time, the precise risk convergence rate of gradient descent for any smooth loss function and in all regimes of T and n.

In some more detail, our results assume the following form. Let $C_{\phi,\beta}$ be the class of nonnegative, convex and β-smooth loss functions $\ell(u)$ that decay to zero (as $u \to \infty$) faster than a reference "tail function" $\phi : [0, \infty) \to \mathbb{R}$. The function ϕ is merely used to quantify how rapidly the tails of loss functions in $C_{\phi,\beta}$ decay to zero. Then, the risk upper and lower bounds we prove for gradient descent are of the following form:[2]

$$\frac{\beta r_{\phi,T}^2}{\gamma^2 T} + \frac{\beta r_{\phi,T}^2}{\gamma^2 n}. \tag{1}$$

---

[2]For simplicity, we specialize the bounds here to gradient descent with stepsize $\eta = \Theta(1/\beta)$, but the bounds hold more generally to any stepsize smaller than $O(1/\beta)$.

Table 2: Examples of tight population risk bounds established in this paper for gradient descent on γ-separable data, instantiated for several different loss tail decay rates. Details on the derivation of the bounds from the general Theorems 1 and 2 can be found in Appendix A.

Tail decay (ϕ)	$r_{\phi,T}$	Risk bound
$\exp(-x)$	$\log(T)$	$\Theta\left(\dfrac{\log^2(T)}{\gamma^2 T} + \dfrac{\log^2(T)}{\gamma^2 n}\right)$
$x^{-\alpha}$	$(\gamma^2 T)^{\frac{1}{\alpha+2}}$	$\Theta\left(\left(\dfrac{1}{\gamma}\right)^{\frac{2\alpha}{2+\alpha}}\left(\dfrac{1}{T^{\frac{\alpha}{2+\alpha}}} + \dfrac{T^{\frac{2}{2+\alpha}}}{n}\right)\right)$
$\exp(-x^\alpha)$	$\log^{\frac{1}{\alpha}}(T)$	$\Theta\left(\dfrac{\log^{\frac{2}{\alpha}}(T)}{\gamma^2 T} + \dfrac{\log^{\frac{2}{\alpha}}(T)}{\gamma^2 n}\right)$

The bounds depend on the tail function ϕ through the term $r_{\phi,T}$ that, roughly, equals $\phi^{-1}(\varepsilon)$ for ε chosen such that $\varepsilon \approx (\phi^{-1}(\varepsilon))^2/\gamma^2 T$ (for the precise statements refer to Theorems 1 and 2, and for concrete examples of implied bounds, see Table 2).

The form of Eq. (1) resembles the bounds given in recent work by Schliserman and Koren [12]. However, the latter work imposed an extraneous "self-boundedness" assumption (stronger than smoothness, see Table 1) that we do not require, and they did not establish the tightness of their bounds, as we do in this paper by providing matching lower bounds. On the flip side, their upper bounds apply in a broader stochastic convex optimization setup, whereas our bounds are specialized to generalized linear models for classification.

In terms of techniques, our proof methodology is also distinctly (and perhaps somewhat surprisingly) simple compared to previous work. Our upper bounds rely on two elementary properties that gradient methods admit when applied to a smooth and realizable objective: low training loss and low norm of the produced solutions. Both properties are obtained from fairly standard arguments and convergence bounds for smooth gradient descent, when paired with conditions implied by the decay rate of the loss function. Finally, to bound the gap between the population risk and the empirical risk, we employ a classical result of Srebro et al. [15] that bounds the generalization gap of linear models in the so-called "low-noise" (i.e., nearly realizable) smooth regime using local Rademacher complexities. Somewhat surprisingly, this simple combination of tools already give sharp risk results that improve upon the state-of-the-art [13, 12, 16] both in terms of tightness of the bounds and the light set of assumptions they rely on.

We remark here that the proof scheme summarized above can be generalized to essentially any gradient method for which one can prove simultaneously bounds on the optimization error and the norm of possible solutions produced by the algorithm. In the sequel, we focus for concreteness on standard gradient descent, but in Appendix B we also give bounds for Stochastic Gradient Descent (SGD), which is shown to admit both of these properties with high probability over the input sample.

An interesting conclusion from our bounds pertains to the significance of early stopping. We see that gradient descent, even when applied on a smooth loss functions that decay rapidly to zero, might overfit with respect the the surrogate loss (i.e., reach a trivial $\Theta(1)$ risk) as the number of steps T grows. The time by which gradient descent starts overfitting depends on the tail decay rate of the loss function: the slower the decay rate, the shorter the time it takes to overfit. Interestingly, a similar phenomenon may not occur for the zero-one loss of the trained predictor. For example, Soudry et al. [14] show that with the logistic loss, gradient descent does not overfit in terms of the zero-one as T approached infinity—whereas, at the same time, it *does* overfit as $T \to \infty$ in terms of the logistic loss itself, as seen from our lower bounds.

Summary of contributions.   To summarize, the main contribution of this paper are as follows:

• Our first main result (in Section 3) is a high probability risk upper bound for gradient descent in the setting of separable classification with a convex and smooth loss function. Our bound matches the best known upper bounds [13, 12] and greatly extend them to allow for virtually any convex and smooth loss function, considerably relaxing various assumptions imposed by prior work.

- Our second main result (in Section 4) is a nearly matching lower bound for the risk of gradient descent, establishing the tightness of our analysis given the smoothness and tail-decay conditions. The tightness of our bounds holds across tail decay rates and different regimes of the parameters T, n and γ. To our knowledge, these constitute the first (let alone tight) risk lower bounds in this context, despite a long line of work on linear classification with separable data.

- We also provide analogous results for Stochastic Gradient Descent with replacement (in Appendix B), mainly to emphasize that that our analysis uses only two elementary properties of the optimization algorithm: low optimization error and low norm of the produced solution. The same analysis can be generalized to any gradient method that admits these two properties.

1.2 Additional related work

Unregularized gradient methods on separable data. The most relevant work to ours is of Schliserman and Koren [12], who used algorithmic stability and two simple conditions of self-boundedness and realizability which the loss functions hold, to get generalization bounds for gradient methods with constant step size in the general setting of stochastic convex and smooth optimization. Then, they derived risk bounds which hold in expectation for the setting of linear classification with separable data for every loss function which decays to 0. The exact bound was depend in the rate of decaying to 0 of the function. For the Lipschitz case their risk bound with respect to the loss function ℓ is $O(\ell^{-1}(\varepsilon)^2 / \gamma^2 T + \ell^{-1}(\varepsilon)^2 / \gamma^2 n)$ for any choice of ε such that $\varepsilon / \ell^{-1}(\varepsilon)^2 \leq 1/\gamma^2 T$. For example, for the logistic loss, the bound translates to $O(\log^2(T)/\gamma^2 T + \log^2(T)/\gamma^2 n)$.

In another work, Telgarsky [16], showed a high probability risk bound for $T \leq n$ for Batch Mirror Decent with step size $\eta \simeq 1/\sqrt{T}$ in linear models, using a reference vector, which when selected properly, can be translated to a risk bound of $O(\ell^{-1}(\varepsilon)/\gamma\sqrt{T})$ for gradient descent applied on the loss function ℓ, and to a $O(\log T/\gamma\sqrt{T})$ for the logistic loss.

Fast rates for smooth and realizable optimization. The problem of smooth and realizable optimization, also known as the "low-noise" regime of stochastic optimization, is a very well researched problem. Srebro et al. [15] showed that stochastic gradient descent achieved risk bound of $O(1/n)$ in this setting. For linear models, they also showed that ERM achieve similar fast rates by using local Rademacher complexities. Later [10] showed that SGD converges linearly when the loss function is also strongly convex. In more recent works, [7] used stability arguments to show that SGD with replacement with $T = n$ achieve risk of $O(1/n)$.

Lower bounds. A lower bound related to ours appears in Ji and Telgarsky [4]. In this work, the authors showed a lower bound of $\|w_t' - w^*\| \geq \log(n)/\log(T)$. In our work, however, we get lower bound directly for the loss itself and not for this objective. More recently, Shamir [13] showed a lower bound of $\Omega(1/\gamma^2 T)$ for the empirical risk of GD when applied on the logistic loss which is tight up to log factors. When generalizing this technique for other objectives, e.g., functions that decay polynomially to zero, the log factors become polynomial factors and the bound becomes not tight, even in the regime $T \ll n$ where the $1/T$ term in the bounds is dominant. In contrast, we establish nearly tight bounds for virtually any tail decay rate, and in all regimes of T and n.

2 Problem Setup

We consider the following typical linear classification setting. Let $\mathcal{D}$ be distribution over pairs (x, y), where $x \in \mathbb{R}^d$ is a d-dimensional feature vector and $y \in \mathbb{R}$ is a real number that represents the corresponding label. We assume that data is scaled so that $\|x\| \leq 1$ and $|y| \leq 1$ (with probability one with respect to $\mathcal{D}$). We focus on the setting of *separable*, or *realizable*, linear classification with margin. Formally, we make the following assumption.

Assumption 1 (realizability)**.** There exists a unit vector $w^* \in \mathbb{R}^d$ and $\gamma > 0$ such that $y(w^* \cdot x) \geq \gamma$ almost surely with respect to the distribution $\mathcal{D}$.

Equivalently, we will identify each pair (x, y) with the vector $z = yx$, and realizability implies that $w^* \cdot z \geq \gamma$ with probability 1. Our assumptions then imply that $\|z\| \leq 1$ with probability 1.

Given a nonnegative loss function $\ell : \mathbb{R} \to \mathbb{R}^+$, the objective is to determine a model $w \in \mathbb{R}^d$ that minimizes the (population) risk, defined as the expected value of the loss function over the distribution

4

$\mathcal{D}$, namely

$$L(w) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(w \cdot z)]. \tag{2}$$

For finding such a model, we use a set of training examples $S = \{z_1, ..., z_n\}$ which drawn i.i.d. from $\mathcal{D}$ and an empirical proxy, the *empirical risk*, which is defined as

$$\widehat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w \cdot z_i). \tag{3}$$

2.1 Loss functions

The loss functions ℓ considered in this paper are nonnegative, convex and β-smooth.[3] We also require that ℓ is strictly monotonically decreasing and $\lim_{u \to \infty} \ell(u) = 0$. The vast majority of loss functions used in supervised learning for classification satisfy these conditions; these include, for example, the logistic loss ($\ell(u) = \log(1 + e^{-u})$), the probit loss ($\ell(u) = -\log(\frac{1}{2} - \frac{1}{2}\operatorname{erf}(u))$), and the squared hinge loss ($\ell(u) = (\max\{1 - u, 0\})^2$).

A main goal of this paper is to quantify how do the achievable bounds on the risk depend on properties of the loss function ℓ used, and most crucially on the rate in which ℓ decays to zero as its argument approaches infinity. To formalize this, we need a couple of definitions.

Definition 1 (tail function). We say that $\phi : [0, \infty) \to \mathbb{R}$ is a *tail function* if ϕ

   (i) is a nonnegative, 1-Lipschitz and β-smooth convex function;
   (ii) is strictly monotonically decreasing such that $\lim_{u \to \infty} \phi(u) = 0$;
   (iii) satisfies $\phi(0) \geq \frac{1}{2}$ and $|\phi'(0)| \geq \frac{1}{2}$.

Every tail function ϕ defines a class of loss functions characterized by the rate ϕ decays to zero.

Definition 2 (ϕ-tailed class). For a tail function ϕ, the class $C_{\phi,\beta}$ is the set of all nonnegative, convex, β-smooth and monotonically decreasing loss functions $\ell : \mathbb{R} \to \mathbb{R}^+$ such that $\ell(u) \leq \phi(u)$ for all $u \geq 0$.

We detail several examples for tail functions in Appendix A.

2.2 Gradient Descent

The algorithm that we focus in this paper is standard gradient descent (GD) with a fixed step size $\eta > 0$ applied to the empirical risk $\widehat{L}$; this method is initialized at $w_1 = 0$ and at each step $t = 1, \ldots, T$ performs an update

$$w_{t+1} = w_t - \eta \nabla \widehat{L}(w_t). \tag{4}$$

The algorithm returns the final model, w_T.

We remark however that most of the results we present in the sequel can be straightforwardly adapted to other gradient methods; we include results for Stochastic Gradient Descent (SGD) in Appendix B, and the same proof techniques can be used to analyze multi-epoch and/or mini-batched SGD, gradient flow, and more.

3 Risk Upper Bounds

We begin by giving a general upper bound for the risk of gradient descent, when the loss function ℓ is taken from the class $C_{\phi,\beta}$. Our main result in this section is the following.

Theorem 1. *Let ϕ be a tail function and let ℓ be any loss function from the class $C_{\phi,\beta}$. Fix T, n and $\delta > 0$. Then, with probability at least $1 - \delta$ (over the random sample S of size n), the output of GD applied on $\widehat{L}$ with step size $\eta \leq 1/2\beta$ initialized at $w_1 = 0$ has*

$$L(w_T) \leq \frac{4K(\phi^{-1}(\varepsilon))^2}{\gamma^2 \eta T} + \frac{32K\beta(\phi^{-1}(\varepsilon))^2 \left(\log^3 n + 4\log\frac{1}{\delta}\right)}{\gamma^2 n} + \frac{4K(\phi^{-1}(\varepsilon))^2 \log\frac{1}{\delta}}{\gamma^2 \eta T n}$$

for any $\varepsilon \leq \frac{1}{2}$ such that $\eta\gamma^2 T \leq (\phi^{-1}(\varepsilon))^2/\varepsilon$, where $K < 10^5$ is a numeric constant.

---

[3] A differentiable function $\ell : \mathbb{R} \to \mathbb{R}$ is said to be β-smooth over $\mathbb{R}$ if $\ell(v) \leq \ell(u) + \ell'(u) \cdot (v - u) + \frac{1}{2}\beta(v - u)^2$ for all $u, v \in \mathbb{R}$.

The expression $(\phi^{-1}(\varepsilon))^2/\varepsilon$ increases indefinitely as ε approaches 0; therefore, for any T, there exists an $\varepsilon > 0$ that satisfies the theorem's condition. For several examples of how this bound is instantiated for different tail decay functions ϕ, see Table 2 and Appendix A.

In the remainder of this section we prove Theorem 1. The structure of the the proof will be as follows: First, we bound the norm of the GD solution; by smoothness and realizability, we get that the norm will remain small compared to a reference point with small loss value. Second, we get a bound on the optimization error in this setting, the relies on the same reference point. Finally, we use a fundamental result due to Srebro et al. [15] (reviewed in the subsection below) together with both bounds to derive the risk guarantee. As discussed broadly in the introduction, this proof scheme can be generalized to other gradient methods which satisfy the properties of model with low norm and low optimization error.

3.1 Preliminaries: Uniform Convergence Using Rademacher Complexity

One property of linear models is that in this class of problems is that we have dimension-independent and algorithm-independent uniform convergence bounds, that enables to bound the difference between the empirical risk and the population risk of a specific model. A main technical tool for bounding this difference is the Rademacher Complexity [2]. The worst-case Rademacher complexity of an hypothesis class H for any sample size n is given by:

$$R_n(H) = \sup_{z_1, \dots z_n} \mathbb{E}_{\sigma \sim \text{Unif}(\{\pm 1\}^n)} \left[\sup_{h \in H} \frac{1}{n} \left| \sum_{i=1}^{n} h(z_i)\sigma_i \right| \right].$$

We are interested in models that achieve low empirical risk on smooth objectives. A fundamental result of Srebro et al. [15] bounds the generalization gap under such conditions:

Proposition 1 (15, Theorem 1)**.** *Let H be a hypothesis class with respect to some non negative and β-smooth function, $\ell(t \cdot y)$, such that for every $w \in H, x, y, |\ell(wx \cdot y)| \leq b$. Then, for any $\delta > 0$ we have, with probability at least $1 - \delta$ over a sample of size n, uniformly for all $h \in H$,*

$$L(h) \leq \widehat{L}(h) + K \left(\sqrt{\widehat{L}(h)} \left(\sqrt{\beta} \log^{1.5}(n) R_n(H) + \sqrt{\frac{b \log \frac{1}{\delta}}{n}} \right) + \beta \log^3(n) R_n^2(H) + \frac{b \log \frac{1}{\delta}}{n} \right).$$

where $K < 10^5$ is a numeric constant.

3.2 Properties of Gradient Descent on Smooth Objectives

In this section we prove that GD satisfies the two desired properties- low norm and low optimization error. We begin with showing that the norm of w_T, the output of GD after T iterations, is low, as stated in the following lemma,

Lemma 1. *Let ϕ be a tail function and let $\ell \in C_{\phi,\beta}$. Fix any $\varepsilon > 0$ and a point $w_\varepsilon^* \in \mathbb{R}^d$ such that $\widehat{L}(w_\varepsilon^*) \leq \varepsilon$ (exists due to realizability). Then, the output of T-steps GD, applied on $\widehat{L}$ with stepsize $\eta \leq 1/\beta$ initialized at $w_1 = 0$ has,*

$$\|w_T\| \leq 2\|w_\varepsilon^*\| + 2\sqrt{\eta \varepsilon T}.$$

Now, we bound the optimization error of GD on every function $\ell \in C_{\phi,\beta}$, by using a variant of Lemma 13 from [12]. The proof is fairly standard and appears in Appendix C.

Lemma 2. *Let ϕ be a tail function and let $\ell \in C_{\phi,\beta}$. Fix any $\varepsilon > 0$ and a point $w_\varepsilon^* \in \mathbb{R}^d$ such that $\widehat{L}(w_\varepsilon^*) \leq \varepsilon$. Then, the output of T-steps GD, applied on $\widehat{L}$ with stepsize $\eta \leq 1/\beta$ initialized at $w_1 = 0$ has,*

$$\widehat{L}(w_T) \leq \frac{\|w_\varepsilon^*\|^2}{\eta T} + 2\varepsilon.$$

3.3 Proof of Theorem 1

We now turn to prove Theorem 1. The proof is a simple consequence of the properties proved above and Proposition 1. We first claim that there exists a model w_ε^* with low norm such that $\widehat{L}(w_\varepsilon^*) \leq \varepsilon$,

which implies, through Lemmas 1 and 2, that w_T of gradient descent has both low optimization error *and* it remains bounded within a ball of small radius. Then, we use Proposition 1 to translate the low optimization error to low risk.

Proof of Theorem 1. First, we show that there exists a model w_ε^* with low norm such that $\widehat{L}(w_\varepsilon^*) \leq \varepsilon$. Let $\varepsilon \leq \frac{1}{2}$ and $\ell \in C_{\phi,\beta}$. By separability, there exists a unit vector w^* such that $w^* \cdot z_i \geq \gamma$ for every z_i in the training set S. Moreover, ℓ is monotonic decreasing. Then, for $w_\varepsilon^* = (\phi^{-1}(\varepsilon)/\gamma)w^*$ and every $z_i \in S$,

$$\ell(w_\varepsilon^* \cdot z_i) = \ell\left(\frac{\phi^{-1}(\varepsilon)}{\gamma}w^* \cdot z_i\right) \leq \ell\left(\frac{\phi^{-1}(\varepsilon)}{\gamma} \cdot \gamma\right) \leq \ell(\phi^{-1}(\varepsilon)).$$

Then, by the fact that $\phi^{-1}(\varepsilon) \geq 0$, we have $\ell(w_\varepsilon^* \cdot z_i) \leq \ell(\phi^{-1}(\varepsilon)) \leq \phi(\phi^{-1}(\varepsilon)) = \varepsilon$ for all i, hence

$$\widehat{L}(w_\varepsilon^*) = \frac{1}{n}\sum_{i=1}^{n}\ell(w_\varepsilon^* \cdot z_i) \leq \varepsilon.$$

Now, for ε such that $\eta\gamma^2 T \leq (\phi^{-1}(\varepsilon))^2/\varepsilon$, we get by Lemma 1,

$$\|w_T\| \leq 2\|w_\varepsilon^*\| + 2\sqrt{\eta\varepsilon T} \leq \frac{4\phi^{-1}(\varepsilon)}{\gamma}.$$

For the same ε, by Lemma 2,

$$\widehat{L}(w_T) \leq \frac{\|w_\varepsilon^*\|^2}{\eta T} + 2\varepsilon \leq 3\frac{\phi^{-1}(\varepsilon)^2}{\gamma^2\eta T}.$$

Denote $B_\varepsilon = \{w : \|w\| \leq r_\varepsilon\}$, where for brevity $r_\varepsilon = 4\phi^{-1}(\varepsilon)/\gamma$. We have, by Lemma 10, $f(x) \leq 2f(y) + \beta\|x - y\|^2$ for all $x, y \in \mathbb{R}^d$ (see proof in Appendix C). Then, together with the fact that $\|z\|, \|z'\| \leq 1$ and choosing ε such that $\varepsilon \leq \phi^{-1}(\varepsilon)^2/\gamma^2\eta T$, with probability 1,

$$b = \max_{w\in B_\varepsilon}|\ell(w \cdot z)| \leq 2\ell(w_\varepsilon^* z) + 4\beta r_\varepsilon^2 \leq 2\varepsilon + 4\beta r_\varepsilon^2 \leq \frac{r_\varepsilon^2}{8\eta T} + 4\beta r_\varepsilon^2.$$

Moreover, B_ε is hypothesis class of linear predictors with norm at most r_ε. We know that the norm of the examples is at most 1, thus, it follows that the Rademacher complexity of B_ε is $R_n(B_\varepsilon) = r_\varepsilon/\sqrt{n}$ [e.g., 6, Theorem 3].

Now, by the choice of ε, we have $\widehat{L}(w_T) \leq 3r_\varepsilon^2/16\eta T$. Thus, Proposition 1 implies that with probability at least $1 - \delta$, for every $w \in B_\varepsilon$ and any ε such that $\varepsilon \leq \phi^{-1}(\varepsilon)^2/\gamma^2\eta T$,

$$L(w_T) \leq \frac{3r_\varepsilon^2}{16\eta T} + K\left(\sqrt{\frac{3r_\varepsilon^2}{16\eta T}}\left(\sqrt{\beta}\log^{1.5}n\frac{r_\varepsilon}{\sqrt{n}} + \sqrt{\frac{b\log\frac{1}{\delta}}{n}}\right) + \beta\log^3 n\frac{r_\varepsilon^2}{n} + \frac{b\log\frac{1}{\delta}}{n}\right).$$

Plugging in the bound on b, dividing by r_ε^2 and using twice the fact that $xy \leq \frac{1}{2}x^2 + \frac{1}{2}y^2$ for all x, y,

$$\frac{L(w_T)}{r_\varepsilon^2} \leq \frac{3}{16\eta T} + K\left(\sqrt{\frac{3}{16\eta T}}\left(\frac{\sqrt{\beta}\log^{1.5}n}{\sqrt{n}} + \sqrt{\frac{(\frac{1}{8\eta T} + 4\beta)\log\frac{1}{\delta}}{n}}\right) + \frac{\beta\log^3 n}{n} + \frac{(\frac{1}{8\eta T} + 4\beta)\log\frac{1}{\delta}}{n}\right)$$

$$\leq \frac{3}{16\eta T} + K\left(\sqrt{\frac{3}{16\eta T}}\left(\frac{\sqrt{\beta}\log^{1.5}n}{\sqrt{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{8\eta Tn} + \frac{4\beta\log\frac{1}{\delta}}{n}}\right) + \frac{\beta\log^3 n}{n} + \frac{\log\frac{1}{\delta}}{8\eta Tn} + \frac{4\beta\log\frac{1}{\delta}}{n}\right)$$

$$\leq \frac{3}{16\eta T} + K\left(\frac{3}{32\eta T} + \frac{1}{2}\left(\frac{\sqrt{\beta}\log^{1.5}n}{\sqrt{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{8\eta Tn} + \frac{4\beta\log\frac{1}{\delta}}{n}}\right)^2 + \frac{\beta\log^3 n}{n} + \frac{\log\frac{1}{\delta}}{8\eta Tn} + \frac{4\beta\log\frac{1}{\delta}}{n}\right)$$

$$\leq \frac{3}{16\eta T} + K\left(\frac{3}{32\eta T} + \frac{\beta\log^3 n}{n} + \frac{\log\frac{1}{\delta}}{8\eta Tn} + \frac{4\beta\log\frac{1}{\delta}}{n} + \frac{\beta\log^3 n}{n} + \frac{\log\frac{1}{\delta}}{8\eta Tn} + \frac{4\beta\log\frac{1}{\delta}}{n}\right)$$

$$\leq \frac{7K}{32\eta T} + \frac{2K\beta\left(\log^3 n + 4\log\frac{1}{\delta}\right)}{n} + \frac{K\log\frac{1}{\delta}}{4\eta Tn}$$

The theorem follows by rearranging the inequality. $\qquad\square$

4   Risk Lower Bounds

In this section we present our second main result: a lower bound showing that the bound we proved in Section 3 for gradient descent is essentially tight for loss functions in the class $C_{\phi,\beta}$, for any given tail function ϕ and any $\beta > 0$. Formally, we prove the following theorem.

Theorem 2. *There exists a constant C such that the following holds. For any tail function ϕ, sample size $n \geq 35$ and any T, there exist a distribution $\mathcal{D}$ and a loss function $\ell \in C_{\phi,\beta}$, such that for T-steps GD over a sample $S = \{z_i\}_{i=1}^n$ sampled i.i.d. from $\mathcal{D}$, initialized at $w_1 = 0$ with stepsize $\eta \leq 1/2\beta$, it holds that*

$$\mathbb{E}[L(w_T)] \geq C\frac{\beta(\phi^{-1}(128\varepsilon))^2}{\gamma^2 n} + C\frac{(\phi^{-1}(8\varepsilon))^2}{\gamma^2 \eta T}.$$

for any $\varepsilon < \frac{1}{256}$ such that $\eta\gamma^2 T \geq (\phi^{-1}(\varepsilon))^2/\varepsilon$.

We remark that the right-hand side of the bound is well defined, as we restrict ε to be sufficiently small so as to ensure that all arguments to ϕ^{-1} are at most $\frac{1}{2}$ (recall that ϕ admits all values in $[0, \frac{1}{2}]$ due to our assumptions that $\phi(0) \geq \frac{1}{2}$). Further, the lower bound above matches the upper bound given in Theorem 1 up to constants, unless the tail function ϕ decays extremely slowly, and slower than any polynomial (at this point, however, the entire bound becomes almost vacuous).

To prove Theorem 2, we consider two different regimes: the first is where $T \gg n$, when the first term in the right-hand side of the bound is dominant; and the $T \ll n$ regime where the second term is dominant. We begin by focusing on the first regime, and prove the following.

Lemma 3. *There exists a constant C_1 such that the following holds. For any tail function ϕ, sample size $n \geq 35$ and any γ and T, there exist a distribution $\mathcal{D}$ with margin γ, a loss function $\ell \in C_{\phi,\beta}$ such that for GD over a sample $S = \{z_i\}_{i=1}^n$ sampled i.i.d. from $\mathcal{D}$, initialized at $w_1 = 0$ with stepsize $\eta \leq 1/2\beta$, it holds that*

$$\mathbb{E}[L(w_T)] \geq C_1\frac{\beta\phi^{-1}(128\varepsilon)^2}{\gamma^2 n},$$

for any $\varepsilon < \frac{1}{256}$ such that $\eta\gamma^2 T \geq (\phi^{-1}(\varepsilon))^2/\varepsilon$.

To prove the lemma, we construct a learning problem for which the risk of GD can be lower bounded. We design a loss function ℓ which exhibits a decay to zero at a rate similar to that of the function ϕ for $x > 0$, and behaves as a quadratic function for $x \leq 0$. The distribution $\mathcal{D}$ is constructed such that there are two fixed examples in its support, denoted z_1 and z_2, that are both sampled with constant probability and are almost opposite in direction to each other: z_2 is aligned with $-z_1$ except for a small component, denoted v, which is orthogonal to z_1. Crucially, since z_1 and z_2 point in nearly opposite directions, minimizing optimization error requires the GD iterate to have a substantial component aligned with the vector v (otherwise, the loss for z_2 will be large). Finally, we define another example, denoted as z_3, that has a significant component in the opposite direction $-v$. We arrange the distribution $\mathcal{D}$ so that z_3 is sampled with with probability roughly $1/n$. The lower bound now follows by observing that, if z_3 is sampled at test time, but did not appear in the training set (this happens with probability $\approx 1/n$), the test error is quadratically large in the magnitude of the GD iterate along $-v$.

We now turn to formally proving Lemma 3.

Proof. Given $\gamma \leq \frac{1}{8}$, let us define the following distribution $\mathcal{D}$:

$$\mathcal{D} = \begin{cases} z_1 := (1, 0, 0) & \text{with prob. } \frac{59}{64}(1 - \frac{1}{n}); \\ z_2 := (-\frac{1}{2}, 3\gamma, 0) & \text{with prob. } \frac{5}{64}(1 - \frac{1}{n}); \\ z_3 := (0, -\frac{1}{8}, 4\gamma + \frac{1}{4}) & \text{with prob. } \frac{1}{n}, \end{cases}$$

and loss function:

$$\ell(x) = \begin{cases} \phi(x) & x \geq 0; \\ \phi(0) + \phi'(0)x + \frac{\beta}{2}x^2 & x < 0. \end{cases}$$

First, note that the distribution is separable: for $w^* = (\gamma, \frac{1}{2}, \frac{1}{4})$ it holds that $w^* z_i = \gamma$ for every $i \in \{1, 2, 3\}$. Moreover, Lemma 11 in Appendix D ensures that indeed $\ell \in C_{\phi,\beta}$.

8

Next, let S be a sample of n i.i.d. examples from $\mathcal{D}$ and let $z' \sim \mathcal{D}$ be a validation example independent from S. Denote by $\delta_2 \in [0, 1]$ the fraction of appearances of z_2 in the sample S, and by A_1, A_2 the following events;

$$A_1 = \{z' = z_3 \wedge z_3 \notin S\}, \qquad A_2 = \left\{\delta_2 \in \left[\tfrac{1}{32}, \tfrac{1}{8}\right]\right\}.$$

In Lemma 12 (found in Appendix D), we show that

$$\Pr(A_1 \cap A_2) \geq \frac{1}{120en}. \tag{5}$$

Furthermore, as in the proof of Theorem 1, there exists a vector w_ε^* which holds $\|w_\varepsilon^*\| \leq \phi^{-1}(\varepsilon)/\gamma$. Then by Lemma 2 and the choice of ε,

$$\widehat{L}(w_T) \leq 2\varepsilon + \frac{2\phi^{-1}(\varepsilon)^2}{\gamma^2} \leq 4\varepsilon. \tag{6}$$

For the remainder of the proof, we condition on the event $A_1 \cap A_2$. First, we show that $w_t \cdot z_2 \geq 0$. Indeed, if it were not the case, then $\ell(w_T \cdot z_2) > \phi(0)$; together with Eq. (6) we obtain

$$\frac{1}{64} > 4\varepsilon \geq \widehat{L}(w_T) \geq \delta_2 \ell(w_T \cdot z_2) \geq \frac{1}{32}\phi(0).$$

which is a contradiction to $\phi(0) \geq \frac{1}{2}$. Moreover, $w_T(1) \geq 0$. Again, we show this by contradiction. Conditioned on A_2, we have $\delta_1 > \frac{7}{8}$. Then, if $w_T(1) < 0$, $\ell(w_T \cdot z_1) > \phi(0)$, and

$$\frac{1}{64} > 4\varepsilon \geq \widehat{L}(w_T) \geq \delta_1 \ell(w_T \cdot z_1) > \frac{7}{8}\phi(0) \geq \frac{7}{16}.$$

which is a contradiction. In addition, we notice that z_3 is the only possible example whose third entry is non zero. Given the event A_1, we know that z_3 is not in S. Equivalently, for every $z \in S$, $z(3) = 0$. As a result,

$$w_t(3) = -\eta \sum_{s=1}^{t-1} \nabla\widehat{L}(w_t) = -\eta \sum_{s=1}^{t-1} \frac{1}{n} \sum_{z \in S} z(3)\ell'(w_s z) = 0.$$

Then, we get that,

$$w_T \cdot z_3 = -\frac{1}{8}w_t(2). \tag{7}$$

Then, using the fact that $w_T \cdot z_2 \geq 0$, $\ell(w_T \cdot z_2) = \phi(w_T \cdot z_2)$, and conditioned on A_2, we have

$$\ell(w_T \cdot z_2) = \phi(w_T \cdot z_2) \leq 32\widehat{L}(w_T)),$$

which implies

$$w_T \cdot z_2 \geq \phi^{-1}(32\widehat{L}(w_T)). \tag{8}$$

Therefore, by combining Eq. (8) with the fact that $w_t(1) \geq 0$,

$$3\gamma w_T(2) \geq -\frac{w_T(1)}{2} + 3\gamma w_T(2) = w_T \cdot z_2 \geq \phi^{-1}(32\widehat{L}(w_T)).$$

which implies, $w_T(2) \geq \frac{1}{3\gamma}\phi^{-1}(32\widehat{L}(w_T))$. By Eq. (7),

$$w_T \cdot z_3 = -\frac{1}{8}w_T(2) \leq -\frac{1}{24\gamma}\phi^{-1}(32\widehat{L}(w_T)).$$

We therefore see that for every ε such that $\varepsilon \geq (\phi^{-1}(\varepsilon))^2/\gamma^2 T\eta$,

$$\ell(w_T \cdot z_3) \geq \frac{\beta}{2}(w_T \cdot z_3)^2 \geq \frac{\beta}{2}\left(\frac{1}{24\gamma}\phi^{-1}(32\widehat{L}(w_T))\right)^2 \geq \frac{\beta}{1152\gamma^2}\left(\phi^{-1}(128\varepsilon)\right)^2,$$

where in the final inequality we again used Eq. (6). We conclude the proof using Eq. (5) and the law of total expectation,

$$\mathbb{E}[L(w_T)] = \mathbb{E}[\ell(w_T \cdot z')] \geq \mathbb{E}[\ell(w_T \cdot z') \mid A_1 \cap A_2]\Pr(A_1 \cap A_2).$$

(Expectations here are taken with respect to both the sample S and the validation example z'.) $\qquad\square$

Second, we focus the second expression in the lower bound of Theorem 2, which is dominant in the early stages of optimization.

Lemma 4. *There exists a constant C_2 such that the following holds. For any tail function ϕ, and for any n, T and γ, there exist a distribution $\mathcal{D}$ with margin γ, a loss function $\ell \in C_{\phi,\beta}$ such that for GD initialized at $w_1 = 0$ with stepsize $\eta \leq 1/2\beta$ over an i.i.d. sample $S = \{z_i\}_{i=1}^{n}$ from $\mathcal{D}$ and $w_1 = 0$ holds*

$$\mathbb{E}[L(w_T)] \geq C_2 \frac{(\phi^{-1}(8\varepsilon))^2}{\gamma^2 T \eta},$$

for any $\varepsilon \leq \frac{1}{16}$ such that $\eta\gamma^2 T \geq \phi^{-1}(\varepsilon)^2/\varepsilon$.

The argument for proving Lemma 4 is similar to that of Lemma 3. Here we define a 1-Lipschitz loss function ℓ that decays to zero at the same rate as ϕ, and a distribution such that there is a possible example z_1 which is sampled with high probability and an almost "opposite" example z_2, that with constant probability, appears rarely in the training set S. The lower bound follows from the fact that although z_2 appears in the dataset, the gradients of the loss function are bounded and thus not large enough so as to make the trained predictor classify z_2 correctly. The full proof can be found in Appendix D.2.

Finally, we derive Theorem 2 directly from Lemmas 3 and 4:

Proof of Theorem 2. Let $C = \frac{1}{2}\min\{C_1, C_2\}$, where C_1 and C_2 are the constants from Lemmas 3 and 4, respectively. If $(\phi^{-1}(8\varepsilon))^2/\gamma^2 T\eta \geq \beta(\phi^{-1}(128\varepsilon))^2/\gamma^2 n$, the theorem follows from Lemma 4; otherwise, it follows from Lemma 3. □

Acknowledgments

References

[1] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

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

[3] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[4] Ziwei Ji and Matus Telgarsky. A refined primal-dual analysis of the implicit bias. *Journal of Environmental Sciences (China) English Ed*, 2019.

[5] Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2109–2136. PMLR, 09–12 Jul 2020.

[6] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[7] 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.

[8] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[9] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR, 2019.

[10] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[11] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[12] Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3380–3394. PMLR, 02–05 Jul 2022.

[13] Ohad Shamir. Gradient methods never overfit on separable data. *The Journal of Machine Learning Research*, 22(1):3847–3866, 2021.

[14] 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.

[15] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[16] Matus Telgarsky. Stochastic linear optimization never overfits with quadratically-bounded losses on general data. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5453–5488. PMLR, 02–05 Jul 2022.

A Derivation of Risk Bounds in Table 2

In this section, we explain how to apply our risk bounds for GD with several popular choices of ϕ-tailed classes.

A.1 Exponentially-tailed functions

We begin with functions that decay exponentially to zero. For any $\alpha > 0$ we define following the tail function, $\phi_\alpha(y) = e^{-y^\alpha}$. There are a lot of popular loss functions loss function $\ell \in C_{\phi_\alpha,\beta}$, e.g., the logistic loss $\ell(y) = \log(1 + e^{-y})$. holds $\ell \in C_{\phi_1,1}$. For this class of functions function, we get the following risk bound:

Corollary 1. *If $\ell \in C_{\phi_\alpha,\beta}$, the output of gradient descent on $\widehat{F}$ with step size $\eta = \frac{1}{2\beta}$ and $w_1 = 0$ holds, with probability of $1 - \delta$, for a suitable choice of ε,*

$$L(w_T) = \tilde{O}\left(\frac{\log^{\frac{2}{\alpha}} T}{\gamma^2 T} + \frac{\log^{\frac{2}{\alpha}} T}{\gamma^2 n}\right).$$

Moreover, if $\gamma^2 T \geq 256^4$ and $\frac{\alpha}{4} \geq \frac{\log\log(\eta\gamma^2 T)}{\log(\eta\gamma^2 T)}$, there exists a function $\bar{\ell} \in C_{\phi_\alpha,\beta}$ and a distribution $\mathcal{D}$ such that

$$\mathbb{E}[\bar{L}(w_T)] = \tilde{\Omega}\left(\frac{\log^{\frac{2}{\alpha}} T}{\gamma^2 T} + \frac{\log^{\frac{2}{\alpha}} T}{\gamma^2 n}\right).$$

Proof. First, $\phi^{-1}(x) = \log^{\frac{1}{\alpha}}(\frac{1}{x})$. For the upper bound we choose $\varepsilon = 1/\eta\gamma^2 T$, which holds $\eta\gamma^2 T \leq (\phi^{-1}(\varepsilon))^2/\varepsilon$. Then,

$$\phi^{-1}(\varepsilon) = \log^{\frac{1}{\alpha}}(\gamma^2\eta T)$$

Then, by Theorem 1, with probability $1 - \delta$,

$$\bar{L}(w_T) \leq \frac{4K(\phi^{-1}(\varepsilon))^2}{\gamma^2\eta T} + \frac{32K\beta(\phi^{-1}(\varepsilon))^2\left(\log^3 n + 4\log\frac{1}{\delta}\right)}{\gamma^2 n} + \frac{4K(\phi^{-1}(\varepsilon))^2\log\frac{1}{\delta}}{\gamma^2\eta Tn}$$

$$= O\left(\frac{(\log\gamma^2 T)^{\frac{2}{\alpha}}}{\gamma^2 T} + \frac{(\log\gamma^2 T)^{\frac{2}{\alpha}}\left(\log^3 n + 4\log\frac{1}{\delta}\right)}{\gamma^2 n}\right)$$

For the lower bound we choose, $\varepsilon = (\gamma^2\eta T)^{-\frac{1}{2}}$, which holds

$$(\eta\gamma^2 T\varepsilon)^{\frac{\alpha}{2}} = (\eta\gamma^2 T)^{\frac{\alpha}{2}}\varepsilon^{\frac{\alpha}{2}} = (\eta\gamma^2 T)^{\frac{\alpha}{2}} \cdot (\eta\gamma^2 T)^{\frac{-\alpha}{4}} = (\eta\gamma^2 T)^{\frac{\alpha}{4}} \geq (\eta\gamma^2 T)^{\frac{\log\log(\eta\gamma^2 T)}{\log(\eta\gamma^2 T)}} = \log(\eta\gamma^2 T)$$

$$\eta\gamma^2 T\varepsilon \geq \log(\eta\gamma^2 T)^{\frac{2}{\alpha}} = \log^{\frac{2}{\alpha}}\frac{1}{\varepsilon^2} = 2^{\frac{2}{\alpha}}\log^{\frac{2}{\alpha}}(\frac{1}{\varepsilon}) \geq \log^{\frac{2}{\alpha}}(\frac{1}{\varepsilon})$$

Then,

$$\phi^{-1}(8\varepsilon) = \log^{\frac{1}{\alpha}}(\frac{1}{8}\gamma\sqrt{\eta T}) = (\frac{1}{2})^{\frac{1}{\alpha}}\log^{\frac{1}{\alpha}}(\frac{1}{64}\gamma^2\eta T) \geq (\frac{1}{2})^{\frac{1}{\alpha}}\log^{\frac{1}{\alpha}}(4\gamma\sqrt{\eta T}) \geq (\frac{1}{2})^{\frac{2}{\alpha}}\log^{\frac{1}{\alpha}}(\gamma^2\eta T)$$

$$\phi^{-1}(128\varepsilon) = \log^{\frac{1}{\alpha}}(\frac{1}{128}\gamma\sqrt{\eta T}) = (\frac{1}{2})^{\frac{1}{\alpha}}\log^{\frac{1}{\alpha}}(\frac{1}{128^2}\gamma^2\eta T) \geq (\frac{1}{2})^{\frac{1}{\alpha}}\log^{\frac{1}{\alpha}}(4\gamma\sqrt{\eta T}) \geq (\frac{1}{2})^{\frac{2}{\alpha}}\log^{\frac{1}{\alpha}}(\gamma^2\eta T)$$

As a result, by Theorem 2, there exists a function $\bar{\ell} \in C_{\phi_\alpha,\beta}$ and a distribution

$$\mathbb{E}[\bar{L}(w_T)] \geq C\frac{\beta(\phi^{-1}(128\varepsilon))^2}{\gamma^2 n} + C\frac{(\phi^{-1}(8\varepsilon))^2}{\gamma^2\eta T}$$

$$= \Omega\left(\frac{\log^{\frac{2}{\alpha}}(\gamma^2 T)}{\gamma^2 n} + \frac{\log^{\frac{2}{\alpha}}(\gamma^2 T)}{\gamma^2 T}\right).$$

$\square$

A.2 Polynomially-tailed losses

Now we turn to discuss loss functions with polynomially-decaying tails. For any $\alpha > 0$ we define following the tail function, $\phi_\alpha(y) = y^{-\alpha}$. Again, we get risk bounds for every loss function $\ell \in C_{\phi_\alpha, \beta}$.

Corollary 2. *If $\ell \in C_{\phi_\alpha, \beta}$, the output of gradient descent on $\widehat{F}$ with step size $\eta = \frac{1}{2\beta}$ and $w_1 = 0$ holds, with probability of $1 - \delta$, for a suitable choice of ε,*

$$L(w_T) = \tilde{O}\left(\left(\frac{1}{\gamma}\right)^{\frac{2\alpha}{2+\alpha}}\left(\frac{1}{T^{\frac{\alpha}{2+\alpha}}} + \frac{T^{\frac{2}{\alpha+2}}}{n}\right)\right).$$

Moreover, if $\gamma^2 T \geq$ there exists a function $\bar{\ell} \in C_{\phi_\alpha, \beta}$ and a distribution $\mathcal{D}$ such that

$$\mathbb{E}[\bar{L}(w_T)] = \Omega\left(\left(\frac{1}{\gamma}\right)^{\frac{2\alpha}{2+\alpha}}\left(\frac{1}{T^{\frac{\alpha}{2+\alpha}}} + \frac{T^{\frac{2}{\alpha+2}}}{n}\right)\right).$$

Proof. First, we note that $\phi^{-1}(x) = x^{-\frac{1}{\alpha}}$. Then, for both upper and lower bound, we choose $\varepsilon = \left(\frac{1}{\eta\gamma^2 T}\right)^{\frac{\alpha}{2+\alpha}}$ which holds $(\phi^{-1}(\varepsilon))^2/\varepsilon = \eta\gamma^2 T$. Now,

$$\phi^{-1}(\varepsilon) = (\eta\gamma^2 T)^{\frac{1}{\alpha+2}}$$

Then, by Theorem 1, with probability $1 - \delta$,

$$\bar{L}(w_T) \leq \frac{4K(\phi^{-1}(\varepsilon))^2}{\gamma^2 \eta T} + \frac{32K\beta(\phi^{-1}(\varepsilon))^2\left(\log^3 n + 4\log\frac{1}{\delta}\right)}{\gamma^2 n} + \frac{4K(\phi^{-1}(\varepsilon))^2 \log\frac{1}{\delta}}{\gamma^2 \eta T n}$$

$$= O\left(\frac{(\gamma^2 T)^{\frac{2}{\alpha+2}}}{\gamma^2 T} + \frac{((\gamma^2 T)^{\frac{2}{\alpha+2}}\left(\log^3 n + 4\log\frac{1}{\delta}\right))}{\gamma^2 n}\right)$$

$$= O\left(\frac{1}{(\gamma^2 T)^{\frac{\alpha}{\alpha+2}}} + \frac{T^{\frac{2}{\alpha+2}}\left(\log^3 n + 4\log\frac{1}{\delta}\right)}{\gamma^{\frac{\alpha}{\alpha+2}} n}\right)$$

Moreover,

$$\phi^{-1}(8\varepsilon) = 8^{\frac{-1}{\alpha}}(\eta\gamma^2 T)^{\frac{1}{\alpha+2}}$$

$$\phi^{-1}(128\varepsilon) = 128^{\frac{-1}{\alpha}}(\eta\gamma^2 T)^{\frac{1}{\alpha+2}}$$

As a result, by Theorem 2, there exists a function $\bar{\ell} \in C_{\phi_\alpha, \beta}$ and a distribution

$$\mathbb{E}[\bar{L}(w_T)] \geq C\frac{\beta(\phi^{-1}(128\varepsilon))^2}{\gamma^2 n} + C\frac{(\phi^{-1}(8\varepsilon))^2}{\gamma^2 \eta T}$$

$$= \Omega\left(\frac{(\gamma^2 T)^{\frac{2}{\alpha+2}}}{\gamma^2 n} + \frac{(\gamma^2 T)^{\frac{2}{\alpha+2}}}{\gamma^2 T}\right).$$

$$= \Omega\left(\frac{1}{(\gamma^2 T)^{\frac{\alpha}{\alpha+2}}} + \frac{T^{\frac{2}{\alpha+2}}}{\gamma^{\frac{\alpha}{\alpha+2}} n}\right)$$

$\square$

B Upper bound for Stochastic Gradient Descent With Replacement

We also show generalization bound for Stochastic Gradient Descent (SGD) which is a randomized algorithm which achieves low optimization error in high probability.

Given a dataset $S = \{z_1, ...z_n\}$, we define $\ell_i(w) = \ell(w \cdot z_i)$. Then, SGD with replacement is initialized at a point $w_1 = 0$ and at each step $t = 1, \ldots, T$, samples randomly an index $i_t \in [n]$ and performs an update

$$w_{t+1} = w_t - \eta\nabla\ell_{i_t}(w_t), \tag{9}$$

where $\eta > 0$ is the step size of the algorithm. We consider a standard variant of SGD that returns the average iterate, namely $\overline{w}_T = \frac{1}{T} \sum_{t=1}^{T} w_t$. We start with showing a deterministic bound on the norm of the iterate of the algorithm. Then, we show high probability bound for the empirical risk $\widehat{L}$. Then, the proof of the risk bound is identical to the proof of the risk bound for GD (see Theorem 1). As a result, we will not show the full proof, but only the bounds on the norm (see Lemma 5) and the empirical risk (see Lemma 8).

Lemma 5. *Let T. Let ϕ be a tail function and let $\ell \in C_{\phi,\beta}$. Assume that for every $\varepsilon > 0$ there exists a point w_ε^* such that for every i, $\ell_i(w_\varepsilon^*) \leq \varepsilon$. Then, the output of SGD with replacement, applied on $\widehat{L}$ with step size $\eta \leq \frac{1}{\beta}$ initialized at $w_1 = 0$ has,*

$$\|\overline{w}_T\| \leq 2\|w_\varepsilon^*\| + 2\sqrt{\eta\varepsilon T}.$$

Proof. By Lemma 9 (see Appendix C), we know that for every w, $\|\nabla \ell_{i_t}(w)\|^2 \leq 2\beta \ell_{i_t}(w)$. Therefore, by using $\eta \leq \frac{1}{\beta}$, for every ε,

$$
\begin{aligned}
\|w_{t+1} - w_\varepsilon^*\|^2 &= \|w_t - \eta \nabla \ell_{i_t}(w_t) - w_\varepsilon^*\|^2 \\
&= \|w_t - w_\varepsilon^*\|^2 - 2\eta \langle w_t - w_\varepsilon^*, \nabla \ell_{i_t} \rangle + \eta^2 \|\nabla \ell_{i_t}(w_t)\|^2 \\
&\leq \|w_t - w_\varepsilon^*\|^2 + 2\eta \ell_{i_t}(w_\varepsilon^*) - 2\eta \ell_{i_t}(w_\varepsilon^*) + 2\beta \eta^2 \ell_{i_t}(w_t) \\
&\leq \|w_t - w_\varepsilon^*\|^2 + 2\eta \ell_{i_t}(w_\varepsilon^*) \\
&\leq \|w_t - w_\varepsilon^*\|^2 + 2\eta\varepsilon.
\end{aligned}
$$

By summing until time T,

$$
\begin{aligned}
\|w_T - w_\varepsilon^*\|^2 &\leq \|w_1 - w_\varepsilon^*\|^2 + 2T\eta\varepsilon \\
&= \|w_\varepsilon^*\|^2 + 2\eta\varepsilon T.
\end{aligned}
$$

By taking a square root, using the fact that $\forall x, y \geq 0$ $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$,

$$\|w_T - w_\varepsilon^*\| \leq \|w_\varepsilon^*\| + 2\sqrt{\eta\varepsilon T}. \tag{10}$$

and by using triangle inequality,

$$\|w_T\| \leq \|w_T - w_\varepsilon^*\| + \|w_\varepsilon^*\| \leq 2\|w_\varepsilon^*\| + 2\sqrt{\eta\varepsilon T}.$$

and finally by another use of traingle inequality,

$$\|\overline{w}_T\| \leq \frac{1}{T} \sum_{t=1}^{T} \|w_T\| \leq 2\|w_\varepsilon^*\| + 2\sqrt{\eta\varepsilon T}.$$

$\square$

Now, we prove the next lemma and bound the regret of SGD with replacement.

Lemma 6. *Let T. Let ϕ be a tail function and let $\ell \in C_{\phi,\beta}$. Then, the iterate of SGD with replacement, applied on $\widehat{L}$ with step size $\eta \leq \frac{1}{2\beta}$ initialized at $w_1 = 0$ has, for every $w \in R^d$,*

$$\frac{1}{T} \sum_{t=1}^{T} \ell(w_t \cdot z_{i_t}) - \frac{2}{T} \sum_{t=1}^{T} \ell(w \cdot z_{i_t}) \leq \frac{\|w\|^2}{2\eta T}.$$

Proof. The proof is almost the same as of [12]. For every w, iteration j and possible i_j, by Lemma 9 and convexity,

$$
\begin{aligned}
\|w_{j+1} - w\|^2 &\leq \|w_j - w\|^2 - 2\eta \langle \nabla \ell_{i_j}(w_j)(w_j - w) \rangle + \eta^2 \|\nabla \ell_{i_j}(w_j)\|^2 \\
&\leq \|w_j - w\|^2 + 2\eta \ell(w \cdot z_{i_j}) - 2\eta \ell(w_j \cdot z_{i_j}) + 2\eta^2 L \ell(w_j \cdot z_{i_j}) \\
&\leq \|w_j - w\|^2 + 2\eta \ell(w \cdot z_{i_j}) - \eta \ell(w_j \cdot z_{i_j}).
\end{aligned}
$$

Taking average on $j = 1...T$, we get

$$\frac{1}{T} \sum_{t=1}^{T} \ell(w_t \cdot z_{i_t}) - \frac{2}{T} \sum_{t=1}^{T} \ell(w \cdot z_{i_t}) \leq \frac{\|w\|^2}{2\eta T}.$$

$\square$

We use the following concentration bound, taken from [1], Lemma 9.

Lemma 7. *Suppose $Z_1, \dots Z_T$ is a sequence such that for every $t \leq T$, $\mathbb{E}\,(Z_t | Z_1 \dots Z_{t-1}) = 0$. and let $Z = \sum_{t=1}^{T} Z_t$. Assume that $|Z_t| \leq b$ for all t, and define $V = \sum_{t=1}^{T} \mathbb{E}\left[Z_t^2 \mid \mathcal{F}_{t-1}\right]$. Then, for any $\delta > 0$ and $\lambda \in [0, 1/b]$, with probability of $1 - \delta$,*

$$Z \leq \lambda V + \frac{1}{\lambda} \log \frac{1}{\delta}.$$

Then, we can get high probability guarantee for SGD with replacement.

Lemma 8. *Let T and $\delta > 0$. Let ϕ be a tail function and let $\ell \in C_{\phi,\beta}$. Assume that for every $\varepsilon > 0$ there exists a point w_ε^* such such that for every i, $\ell_i(w_\varepsilon^*) \leq \varepsilon$. Then, the output of SGD with replacement, applied on $\widehat{L}$ with step size $\eta \leq \frac{1}{\beta}$ initialized at $w_1 = 0$ has, with probability of $1 - \delta$,*

$$\widehat{L}\,(\bar{w}_T) \leq \frac{\|w_\varepsilon^*\|^2}{\eta T} + 5\varepsilon + \frac{8\left(3\varepsilon + 16\beta\|w_\varepsilon^*\|^2 + 16\eta\varepsilon T\right)}{T} \log\left(\frac{1}{\delta}\right).$$

Proof. Let w. We define

$$Z_t = \ell(w_\varepsilon^* \cdot z_{i_t}) - \widehat{L}(w_\varepsilon^*) - \ell(w_t \cdot z_{i_t}) + \widehat{L}(w_t).$$

By the fact that w_t, i_t is independent, $\mathbb{E}\,(Z_t | Z_1 \dots Z_{t-1}) = 0$. First,

$$b = \max_t |Z_t|$$

$$\leq \varepsilon + \max_{t,i} \ell(w_t \cdot z_i) \qquad (\ell(w_\varepsilon^* \cdot z_{i_t}) \leq \varepsilon, \text{ definition of } \widehat{L}, \text{ nonnegativity})$$

$$\leq \varepsilon + \max_{i, \|w\| \leq 2\|w_\varepsilon^*\| + 2\sqrt{\eta\varepsilon T}} \ell(w \cdot z_i) \qquad \text{(Lemma 5)}$$

$$\leq \varepsilon + 2\max_i \ell(w_\varepsilon^* \cdot z_i) + 16\beta\|w_\varepsilon^*\|^2 + 16\eta\varepsilon T \qquad \text{(Lemma 10)}$$

$$\leq 3\varepsilon + 16\beta\|w_\varepsilon^*\|^2 + 16\eta\varepsilon T.$$

We denote $b = 3\varepsilon + 16\beta\|w_\varepsilon^*\|^2 + 16\eta\varepsilon T$. Then,

$$\max_t |Z_t| \leq 3\varepsilon + 16\beta\|w_\varepsilon^*\|^2 + 16\eta\varepsilon T = b.$$

Moreover, we denote $\mathbb{E}_t[\cdot]$ the expectation conditioned on the randomness of the algorithm before step t. Then,

$$V = \sum_{t=1}^{T} \mathbb{E}_t\left[\left(\ell(w_t \cdot z_{i_t}) - \widehat{L}(w_t) - \ell(w \cdot z_{i_t}) + \widehat{L}(w)\right)^2\right]$$

$$\leq b \sum_{i=1}^{T} \mathbb{E}_t\left[\left|\ell(w_t \cdot z_{i_t}) - \widehat{L}(w_t) - \ell(w \cdot z_{i_t}) + \widehat{L}(w)\right|\right]$$

$$\leq b \sum_{i=1}^{T} \mathbb{E}_t\left[\ell(w_t \cdot z_{i_t})|\right] + b \sum_{i=1}^{T} \mathbb{E}_t\left[\ell(w \cdot z_{i_t})|\right] + b \sum_{i=1}^{T} \mathbb{E}_t\left[\widehat{L}(w_t)|\right] + b \sum_{i=1}^{T} \mathbb{E}_t\left[\widehat{L}(w)|\right]$$

$$\leq 2b \sum_{i=1}^{T} \widehat{L}(w_t) + 2bT \sum_{i=1}^{T} \widehat{L}(w)$$

By Lemma 7, with probability $1 - \delta$ by choosing $\lambda = \frac{1}{4b}$,

$$\sum_{t=1}^{T} Z_t \leq \frac{1}{2} \sum_{i=1}^{T} \widehat{L}(w_t) + \frac{1}{2}T\widehat{L}(w) + \frac{4b}{T} \log\left(\frac{1}{\delta}\right).$$

Then, with probability $1 - \delta$,

$$\frac{1}{T} \sum_{i=1}^{T} \widehat{L}(w_t) - \widehat{L}(w) = \frac{1}{T} \sum_{t=1}^{T} \ell(w_t \cdot z_{i_t}) - \ell(w \cdot z_{i_t}) + \frac{1}{T} \sum_{t=1}^{T} Z_t$$

$$\leq \frac{1}{T} \sum_{i=1}^{T} \ell(w_t \cdot z_{i_t}) - \frac{1}{T} \sum_{i=1}^{T} \ell(w \cdot z_{i_t}) + \frac{1}{2T} \sum_{t=1}^{T} \widehat{L}(w_t) + \frac{1}{2}\widehat{L}(w) + 4b \log\left(\frac{1}{\delta}\right).$$

15

By organizing,

$$\frac{1}{2T}\sum_{i=1}^{T}\widehat{L}(w_t) \le \frac{1}{T}\sum_{i=1}^{T}\ell(w_t \cdot z_{i_t}) - \frac{1}{T}\sum_{i=1}^{T}\ell(w \cdot z_{i_t}) + \frac{3}{2}\widehat{L}(w) + \frac{4b}{T}\log\left(\frac{1}{\delta}\right).$$

Moreover, with probability $1 - \delta$, by Lemma 6 and Jensen inequality

$$\widehat{L}\left(\frac{1}{T}\sum_{i=1}^{T}w_t\right) \le \frac{1}{T}\sum_{i=1}^{T}\widehat{L}(w_t) \le \frac{2}{T}\sum_{i=1}^{T}\ell(w_t \cdot z_{i_t}) - \frac{2}{T}\sum_{i=1}^{T}\ell(w \cdot z_{i_t}) + 3\widehat{L}(w) + \frac{8b}{T}\log\left(\frac{1}{\delta}\right)$$

$$= 2\left(\frac{1}{T}\sum_{i=1}^{T}\ell(w_t \cdot z_{i_t}) - \frac{2}{T}\sum_{i=1}^{T}\ell(w \cdot z_{i_t})\right) + 2\widehat{L}(w) + 3\widehat{L}(w) + \frac{8b}{T}\log\left(\frac{1}{\delta}\right)$$

$$\le \frac{\|w\|^2}{\eta T} + 5\widehat{L}(w) + \frac{8b}{T}\log\left(\frac{1}{\delta}\right).$$

Finally, for $w = w_\varepsilon^*$,

$$\widehat{L}(\bar{w}_T) \le \frac{\|w_\varepsilon^*\|^2}{\eta T} + 5\varepsilon + \frac{8\left(3\varepsilon + 16\beta\|w_\varepsilon^*\|^2 + 16\eta\varepsilon T\right)}{T}\log\left(\frac{1}{\delta}\right).$$

$\square$

C Proofs of Section 3

Proof of Lemma 1. From β-smoothness, we know that $\|\nabla\widehat{L}(w)\|^2 \le 2\beta\widehat{L}(w)$ for any w (see Lemma 9 in Appendix C). Therefore, by using $\eta \le 1/\beta$, for every ε,

$$\|w_{t+1} - w_\varepsilon^*\|^2 = \|w_t - \eta\nabla\widehat{L}(w_t) - w_\varepsilon^*\|^2$$

$$= \|w_t - w_\varepsilon^*\|^2 - 2\eta\langle w_t - w_\varepsilon^*, \nabla\widehat{L}(w_t)\rangle + \eta^2\|\nabla\widehat{L}(w_t)\|^2$$

$$\le \|w_t - w_\varepsilon^*\|^2 + 2\eta\widehat{L}(w_\varepsilon^*) - 2\eta\widehat{L}(w_t) + 2\beta\eta^2\widehat{L}(w_t)$$

$$\le \|w_t - w_\varepsilon^*\|^2 + 2\eta\widehat{L}(w_\varepsilon^*)$$

$$\le \|w_t - w_\varepsilon^*\|^2 + 2\eta\varepsilon.$$

By summing until time T,

$$\|w_T - w_\varepsilon^*\|^2 \le \|w_1 - w_\varepsilon^*\|^2 + 2T\eta\varepsilon = \|w_\varepsilon^*\|^2 + 2\eta\varepsilon T.$$

By taking a square root, using the fact that $\forall x, y \ge 0$ $\sqrt{x + y} \le \sqrt{x} + \sqrt{y}$ and using triangle inequality,

$$\|w_T\| \le \|w_T - w_\varepsilon^*\| + \|w_\varepsilon^*\| \le 2\|w_\varepsilon^*\| + 2\sqrt{\eta\varepsilon T}.$$

$\square$

We rely on the following standard lemma about smooth functions (proof can be found in, e.g., 11, or in 15).

Lemma 9. *For a non-negative and β-smooth $f : \mathbb{R}^d \to \mathbb{R}$, it holds that $\|\nabla f(w)\|^2 \le 2\beta f(w)$ for all $w \in \mathbb{R}^d$.*

Proof of Lemma 2. The proof follows the argument of [12]. First, by β-smoothness, for every t and $\eta \le 1/\beta$,

$$\widehat{L}(w_{t+1}) \le \widehat{L}(w_t) + \nabla\widehat{L}(w_t) \cdot (w_{t+1} - w_t) + \frac{\beta}{2}\|w_{t+1} - w_t\|^2$$

$$= \widehat{L}(w_t) - \eta\|\nabla\widehat{L}(w_t)\|^2 + \frac{\eta^2\beta}{2}\|\nabla\widehat{L}(w_t)\|^2$$

$$\le \widehat{L}(w_t) - \frac{\eta}{2}\|\nabla\widehat{L}(w_t)\|^2$$

$$\le \widehat{L}(w_t).$$

16

Hence,

$$\widehat{L}(w_T) \le \frac{1}{T} \sum_{t=1}^{T} \widehat{L}(w_t). \tag{11}$$

Moreover, from standard regret bounds for gradient updates, for any $w \in \mathbb{R}^d$,

$$\frac{1}{T} \sum_{t=1}^{T} (\widehat{L}(w_t) - \widehat{L}(w)) \le \frac{\|w_1 - w\|^2}{2\eta T} + \frac{\eta}{2T} \sum_{t=1}^{T} \|\nabla \widehat{L}(w_t)\|^2.$$

By Lemma 9,

$$\frac{1}{T} \sum_{t=1}^{T} (\widehat{L}(w_t) - \widehat{L}(w)) \le \frac{\|w\|^2}{2\eta T} + \frac{\eta \beta}{T} \sum_{t=1}^{T} \widehat{L}(w_t).$$

Using $\eta \le 1/2\beta$ gives

$$\frac{1}{T} \sum_{t=1}^{T} \widehat{L}(w_t) \le \frac{\|w\|^2}{\eta T} + 2\widehat{L}(w).$$

For $w = w_\varepsilon^*$ we get by Eq. (11),

$$\widehat{L}(w_T) \le \frac{1}{T} \sum_{t=1}^{T} \widehat{L}(w_t) \le \frac{\|w_\varepsilon^*\|^2}{\eta T} + 2\widehat{L}(w_\varepsilon^*) \le \frac{\|w_\varepsilon^*\|^2}{\eta T} + 2\varepsilon.$$

$\square$

Lemma 10. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth and nonnegative function. Then $f(x) \le 2f(y) + \beta\|x - y\|^2$ for all $x, y \in \mathbb{R}^d$.*

Proof. For any $x, y \in \mathbb{R}^d$:

$$f(x) \le f(y) + \nabla f(y) \cdot (x - y) + \frac{\beta}{2}\|x - y\|^2 \qquad (\beta\text{-smoothness})$$

$$\le f(y) + \frac{1}{2\beta}\|\nabla f(y)\|^2 + \frac{\beta}{2}\|x - y\|^2 + \frac{\beta}{2}\|x - y\|^2 \qquad (\forall c > 0 : xy \le \tfrac{1}{2c}x^2 + \tfrac{c}{2}y^2)$$

$$\le 2f(y) + \beta\|x - y\|^2. \qquad (\text{Lemma 9})$$

$\square$

D   Proof of Section 4

D.1   Proof of Lemma 3

Lemma 11. *Let ϕ be a tail function. Let $\ell(x)$ be the following function,*

$$\ell(x) = \begin{cases} \phi(x) & \text{if } x \ge 0; \\ \phi(0) + x\phi'(0) + \frac{\beta}{2}x^2 & \text{if } x < 0. \end{cases}$$

Then, $\ell \in C_{\phi,\beta}$.

Proof. First, it is easy to verify that ℓ is continuously differentiable. Second, ℓ is non negative: for $x \ge 0$ by the non negativity of ϕ and for $x < 0$ by the fact that $\phi'(0) \le 0$. Moreover, ℓ is convex. We need to prove that every $x < y$, $\ell'(x) \le \ell'(y)$ For $x, y < 0$, we get it by the convexity of ϕ. For $x, y > 0$, we get it by the fact ℓ there is a sum of convex function and linear function. For $x < 0 < y$, by the convexity of ϕ,

$$\ell'(x) = \phi'(0) + \beta x \le \phi'(0) \le \phi'(y).$$

In addition, ℓ is β-smooth. We need to prove that every $x < y$, $\ell'(y) - \ell'(x) \le \beta(y - x)$ For $x, y \ge 0$, we get it by the smoothness of ϕ. For $x, y \le 0$, we get it by the fact that ℓ is a sum of β-smooth function and a linear function. For $x \le 0 \le y$, by the smoothness of ϕ,

$$\ell'(y) - \ell'(x) = \phi'(y) - \phi'(0) - \beta x \le \beta(y - x).$$

17

Finally, ℓ is strictly monotonically decreasing. We need to prove that every $x < y$, $\ell(y) > \ell(x)$. For $x, y > 0$, we get it by the monotonicity of ϕ. For $x < y < 0$,

$$\ell(y) = \phi(0) + \phi'(0)y + \frac{\beta}{2}y^2 \leq \phi(0) + \phi'(0)x + \frac{\beta}{2}x^2 = \ell(x).$$

For $x < 0 < y$,

$$\ell(y) = \phi(y) \leq \phi(0) \leq \phi(0) + \phi'(0)x + \frac{\beta}{2}x^2 = \ell(x).$$

$\square$

Lemma 12. *Let $S \sim \mathcal{D}^n$ be a sample of size n, and let $z' \sim \mathcal{D}$ be a validation example. Moreover, Assume $n \geq 35$ and let δ_2 be the fraction of z_2 in S. We define the following event,*

$$A = \{z_3 \notin S\} \cap \{z' = z_3\} \cap \{\delta_2 \in [\tfrac{1}{32}, \tfrac{1}{8}]\}.$$

Then,

$$\Pr(A) \geq \frac{1}{120en}$$

Proof. The proof follows directly by Lemma 13 and Lemma 14. We define the following events,

$$A_1 = \{z_3 \notin S\} \cap \{z' = z_3\}, A_2 = \{\delta_2 \in [\tfrac{1}{32}, \tfrac{1}{8}]\}.$$

By Lemma 13.

$$\Pr(A_1) \geq \frac{1}{2en}.$$

By *Lemma* 14,

$$\Pr(A_2|A_1) \geq \frac{1}{60}.$$

Then, combining both results,

$$\Pr(A) \geq \Pr(A_1)\Pr(A_2 \mid A_1) \geq \frac{1}{120en}.$$

$\square$

Lemma 13. *Let $S \sim \mathcal{D}^n$ be a sample of size n, and let $z' \sim \mathcal{D}$ be a validation example. Then,*

$$\Pr(A_1) = \Pr(z_3 \notin S \wedge z' = z_3) \geq \frac{1}{2en}.$$

Proof. First, we know that,

$$\Pr(z_3 \notin S \wedge z' = z_3) = \Pr(z' = z_3) \cdot \Pr(z' \notin S \mid z' = z_3).$$

and,

$$\Pr(z' = z_3) = \frac{1}{n}.$$

Moreover,

$$\Pr(z' \notin S \mid z' = z_3) = P(z_3 \notin S) = (1 - \frac{1}{n})^n \geq \frac{1}{e}(1 - \frac{1}{n}) \geq \frac{1}{2e}.$$

Combining everything together proves our claim. $\square$

Lemma 14. *Assume $n \geq 35$ and let δ_2 be the fraction of z_2 in S. Then*

$$\Pr(A_2 \mid A_1) = \Pr(\delta_2 \in [\tfrac{1}{32}, \tfrac{1}{8}] \mid A_1) \geq \frac{1}{60}.$$

18

Proof. Let $p_i' = \Pr(z_i = z_2 \mid A_1)$. Note that by the fact that $i \neq j$, z_i, z_j are indepndent. Then, for every $i \neq j$, $p_i' = p_j'$. Then, by the fact that every example is independent,

$$
\begin{aligned}
p_i' &= \Pr(z_i = z_2 \mid z_3 \notin S) \\
&= \Pr(z_i = z_2 \mid z_i \neq z_3) \\
&= \frac{\Pr(z_i = z_2)}{\Pr(z_i \neq z_3)} \\
&= \frac{1}{1 - \frac{1}{n}} \Pr(z_i = z_2) \\
&= \frac{5}{64}.
\end{aligned}
$$

Then,

$$
\mathbb{E}[\delta_2 \mid A_1] = \frac{1}{n} \sum_{i=1}^{n} \Pr(z_i = z_2 \mid A_1) = \frac{1}{n} \sum_{i=1}^{n} p_i' = \frac{5}{64},
$$

and,

$$
\mathrm{Var}(\delta_2 \mid A_1) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} 1_{\{z_i = z_2\}} \mid A_1\right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(1_{\{z_i = z_2\}} \mid A_1) = \frac{5 \cdot 59}{64^2 n}.
$$

Finally, by Chebyshev's inequality, for $n \geq 35$,

$$
\begin{aligned}
\Pr(A_2 \mid A_1) &= \Pr\left(\delta_2 \in \left[\tfrac{1}{32}, \tfrac{1}{8}\right] \mid A_1\right) \\
&= \Pr\left(\left|\delta_2 - \tfrac{5}{64}\right| \leq \tfrac{3}{64} \mid A_1\right) \\
&= 1 - \Pr\left(\left|\delta_2 - \tfrac{5}{64}\right| \geq \tfrac{3}{64} \mid A_1\right) \\
&\geq 1 - \frac{64^2}{9} \mathrm{Var}(\delta_2 \mid A_1) \\
&= 1 - \frac{5 \cdot 59}{9n} \\
&\geq 1 - \frac{5 \cdot 59}{315} \\
&\geq \frac{1}{60}.
\end{aligned}
$$

$\square$

D.2 Proof of Lemma 4

Lemma 15. *Let ϕ be a tail function. Let $\ell(x)$ be the following function,*

$$
\ell(x) = \begin{cases} \phi(x) & \text{if } x \geq 0; \\ \phi(0) + \phi'(0)x & \text{if } x < 0. \end{cases}
$$

Then, $\ell \in C_{\phi,\beta}$.

Proof. First, it is easy to verify that ℓ is continuously differentiable. Second, ℓ is non negative: for $x \geq 0$ by the non negativity of ϕ and for $x < 0$ by the fact that $\phi'(0) \leq 0$. Moreover, ℓ is convex. We need to prove that every $x < y$, $\ell'(x) \leq \ell'(y)$ For $x, y < 0$, we get it by the convexity of ϕ. For $x, y > 0$, we get it by the linearity of ℓ. For $x < 0 < y$, by the convexity of ϕ,

$$
\ell'(x) = \phi'(0) \leq \phi'(y) = \ell'(y).
$$

In addition, ℓ is β-smooth. We need to prove that every $x < y$, $\ell'(y) - \ell'(x) \leq \beta(y - x)$ For $x, y \geq 0$, we get it by the smoothness of ϕ. For $x, y \leq 0$, we get it by the linearity of ℓ. For $x \leq 0 \leq y$, by the smoothness of ϕ,

$$
\ell'(y) - \ell'(x) = \phi'(y) - \phi'(0) \leq \beta y \leq \beta(y - x).
$$

Finally, ℓ is strictly monotonically decreasing. We need to prove that every $x < y$, $\ell(y) > \ell(x)$. For $x, y > 0$, we get it by the monotonicity of ϕ. For $x < y < 0$,

$$\ell(y) = \phi(0) + \phi'(0)y \le \phi(0) + \phi'(0)x = \ell(x).$$

For $x < 0 < y$,

$$\ell(y) = \phi(y) \le \phi(0) \le \phi(0) + \phi'(0)x = \ell(x).$$

$\square$

Proof of Lemma 4. First, we give a sketch of the proof: The proof argument is similar to that of Lemma 3. A key difference is that the example that GD classifies incorrectly does appear in the dataset (though rarely). We define a 1-Lipschitz loss function ℓ that decays to zero at the same rate as ϕ, and a distribution such that there is another possible example z_1 and an almost "opposite" example z_2, that with constant probability, appears limited times in the training samples S. The lower bound follows from the fact that although z_2 appears in the dataset, the gradients of the loss function are sufficiently large so as to make the trained predictor correct on z_2.

For the proof, given $\gamma \le \frac{1}{8}$ and $\varepsilon \le \frac{1}{16}$, consider the following distribution;

$$\mathcal{D} = \begin{cases} z_1 := (1, 0) & \text{with prob. } 1 - p; \\ z_2 := (-\frac{1}{2}, 3\gamma) & \text{with prob. } p, \end{cases}$$

where $p = \frac{\phi^{-1}(8\varepsilon)}{72\gamma^2 T \eta}$. Note that the distribution is separable, as for $w^* = (\gamma, \frac{1}{2})$ it holds that $w^* z_1 = w^* z_2 = \gamma$. Further, consider the following loss function;

$$\ell(x) = \begin{cases} \phi(x) & \text{if } x \ge 0; \\ \phi'(0)x + \phi(0) & \text{otherwise.} \end{cases}$$

First, Lemma 15 below ensures that $\ell \in C_{\phi, \beta}$.

Next, let S be a sample of n i.i.d. examples from $\mathcal{D}$. Denote by $\delta_2 \in [0, 1]$ the fraction of appearances of z_2 in the sample S, and by A_1 the event that $\delta_2 \le 2p$. By Markov's inequality, we have $\Pr(A_1) \ge \frac{1}{2}$. Furthermore, as in the proof of Theorem 1, there exists a vector w_ε^* which holds $\|w_\varepsilon^*\| \le \frac{\phi^{-1}(\varepsilon)}{\gamma}$. Then by Lemma 2 and the choice of ε,

$$\widehat{L}(w_T) \le 2\varepsilon + \frac{2\phi^{-1}(\varepsilon)^2}{\gamma^2} \le 4\varepsilon. \tag{12}$$

Now, we assume that A_1 holds. We know that

$$\delta_2 \le 2p \le \frac{\phi^{-1}(8\varepsilon)}{36\gamma^2 T \eta} \le \varepsilon \le \frac{1}{2}, \tag{13}$$

thus, conditioned on A_1 and by Eq. (12),

$$4\varepsilon \ge \widehat{L}(w_T) > (1 - \delta_2)\ell(w_T \cdot z_1) \ge \frac{1}{2}\ell(w_T(1)). \tag{14}$$

If $w_T(1) < 0$, we get that

$$4\varepsilon > \frac{1}{2}\ell(0) = \frac{1}{2}\phi(0) \ge \frac{1}{4}$$

which is a contradiction to our assumption that $\varepsilon \le \frac{1}{16}$. Then $w_T(1) \ge 0$ and from Eq. (14), $8\varepsilon \ge \ell(w_T(1)) = \phi(w_T(1))$. which implies that

$$w_T(1) \ge \phi^{-1}(8\varepsilon). \tag{15}$$

Now, by the fact that $\phi'(0) \le 1$ it follows that ℓ is 1-Lipschitz. Then, from the GD update rule,

$$w_{t+1}(2) = w_t(2) - 3\eta \cdot \gamma \delta_2 \ell'(w_t \cdot z_2) \le w_t(2) + 3\gamma \delta_2 \eta,$$

from which it follows that

$$w_T(2) \le 3\gamma \delta_2 \eta T. \tag{16}$$

From Eqs. (13), (15) and (16) we now obtain that

$$w_T \cdot z_2 \le 9\gamma^2 \delta_2 T\eta - \frac{1}{2}\phi^{-1}(8\varepsilon)$$

$$\le 9\gamma^2 T\eta \frac{\phi^{-1}(8\varepsilon)}{36\gamma^2 T\eta} - \frac{1}{2}\phi^{-1}(8\varepsilon)$$

$$= -\frac{1}{4}\phi^{-1}(8\varepsilon).$$

By the fact that $\forall x < 0 : \ell(x) \ge -\frac{1}{2}x$, this implies that in the event A_1 it holds that:

$$\ell(w_T \cdot z_2) \ge -\frac{1}{2}w_T \cdot z_2 \ge \frac{1}{8}\phi^{-1}(8\varepsilon). \tag{17}$$

Finally, for a new validation example $z' \sim \mathcal{D}$ (independent from the sample S),

$$\Pr(\{z' = z_2\} \cap A_1) = \Pr(z' = z_2 \mid A_1)\Pr(A_1) \ge \frac{1}{2}P(z' = z_2) = \frac{1}{2}p \ge \frac{\phi^{-1}(8\varepsilon)}{144\gamma^2 T\eta}. \tag{18}$$

To conclude, from Eqs. (17) and (18) we have

$$\mathbb{E}[\ell(w_T z')] \ge \mathbb{E}[\ell(w_T z') \mid \{z' = z_2\} \cap A_1]\Pr(\{z' = z_2\} \cap A_1)$$

$$\ge \frac{\phi^{-1}(8\varepsilon)}{144\gamma^2 T\eta} \cdot \frac{1}{8}\phi^{-1}(8\varepsilon)$$

$$= \frac{\phi^{-1}(8\varepsilon)^2}{1152\gamma^2 T\eta}.$$

$\square$