Gradient Methods Never Overfit On Separable Data

Ohad Shamir   OHAD.SHAMIR@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
Rehovot, Israel

Abstract

A line of recent works established that when training linear predictors over separable data, using gradient methods and exponentially-tailed losses, the predictors asymptotically converge in direction to the max-margin predictor. As a consequence, the predictors asymptotically do not overfit. However, this does not address the question of whether overfitting might occur non-asymptotically, after some bounded number of iterations. In this paper, we formally show that standard gradient methods (in particular, gradient flow, gradient descent and stochastic gradient descent) *never* overfit on separable data: If we run these methods for T iterations on a dataset of size m, both the empirical risk and the generalization error decrease at an essentially optimal rate of $\tilde{\mathcal{O}}(1/\gamma^2 T)$ up till $T \approx m$, at which point the generalization error remains fixed at an essentially optimal level of $\tilde{\mathcal{O}}(1/\gamma^2 m)$ regardless of how large T is. Along the way, we present non-asymptotic bounds on the number of margin violations over the dataset, and prove their tightness.
Keywords: Gradient Methods, Margin-based Generalization Bounds, Implicit Bias, Linear Predictors, Linearly Separable Data

1. Introduction

Motivated by empirical observations in the context of neural networks, there is considerable interest nowadays in studying the *implicit bias* of learning algorithms. This refers to the fact that even without any explicit regularization or other techniques to avoid overfitting, the dynamics of the learning algorithm itself biases its output towards "simple" predictors that generalize well.

In this paper, we consider the implicit bias in a well-known and simple setting, namely learning linear predictors ($\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$) for binary classification with respect to linearly-separable data. In a recent line of works (Soudry et al., 2018; Ji and Telgarsky, 2018b; Nacson et al., 2019a; Ji and Telgarsky, 2019b; Dudik et al., 2020), it was shown that if we attempt to do this by minimizing the empirical risk (average loss) over a dataset, using gradient descent and any exponentially-tailed loss (such as the logistic loss), then the predictor asymptotically converges in direction to the max-margin predictor with respect to the Euclidean norm[1]. Since there are standard generalization bounds for predictors which achieve a large margin over the dataset, we get that asymptotically, gradient descent does not overfit, even if we just run it on the empirical risk function without any explicit regularization, and even if the number of iterations T diverges to infinity. In follow-up works,

---

1. Namely, $\arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \min_i \mathbf{x}_i^\top \mathbf{w}$ for a given dataset $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$.

similar results were also obtained for other gradient methods such as stochastic gradient descent and mirror descent (Nacson et al., 2019b; Gunasekar, 2018), and for more complicated predictors such as linear networks, shallow ReLU networks, and linear convolutional networks (Ji and Telgarsky, 2018a, 2019a; Gunasekar et al., 2018).

However, in practice the number of iterations T is some bounded finite number. Thus, the asymptotic results above leave open the possibility that for a wide range of values of T, gradient methods do not achieve a good margin, and possibly overfit. Admittedly, many of these papers do provide finite-time guarantees for linear predictors, which all tend to have the following form: After T iterations, the output of gradient descent $\bar{\mathbf{w}}_T$ (normalized to have unit norm, and with bounded step sizes) generally satisfies[2]

$$\|\bar{\mathbf{w}}_T - \mathbf{w}^*\| \;=\; \mathcal{O}\left(\frac{1}{\log(T)}\right), \tag{1}$$

where $\mathbf{w}^*$ is the max-margin predictor, and the $\mathcal{O}(\cdot)$ notation hides dependencies in the dataset size and the margin attained by $\mathbf{w}^*$. However, such bounds do not satisfactorily address the problem above, since they decay extremely slowly with T. For example, suppose we ask how many iterations T are needed, till we get a predictor which achieves some positive margin on all the data points, assuming there exists a unit-norm predictor $\mathbf{w}^*$ achieving a margin of γ (namely $\min_i \mathbf{x}_i^\top \mathbf{w}^* \geq \gamma$). If all we know is the bound in Eq. (1), we must require $\|\bar{\mathbf{w}}_T - \mathbf{w}^*\| \leq \gamma$, which holds if $T > \exp(\Omega(1/\gamma))$ (and in fact, the actual required bound is much larger due to the hidden dependencies in the $\mathcal{O}(\cdot)$ notation). For realistically small values of γ, this bound on T is unacceptably large. Could it be that gradient methods do not overfit only after so many iterations? We note that in Ji and Telgarsky (2019b), it is shown that Eq. (1) is essentially tight, but this does not preclude the possibility that $\bar{\mathbf{w}}_T$ does not overfit even before getting very close to $\mathbf{w}^*$.

In this paper, we show that this is indeed the case, and in fact, for *any* number of iterations T, gradient methods do not overfit and essentially behave in the best manner we can hope for. Specifically, if the underlying data distribution is separable with margin γ, and we attempt to minimize the average of an exponential or logistic loss over a training set of size m, using standard gradient methods (gradient flow, gradient descent, or stochastic gradient descent), then the generalization error of the resulting predictor (with respect to the $0-1$ loss) is at most $\tilde{\mathcal{O}}(1/\gamma^2 T + 1/\gamma^2 m)$, up to constants and logarithmic factors. For $T \leq m$, this bound is $\tilde{\mathcal{O}}(1/\gamma^2 T)$, which is essentially the same as the optimal upper bound on the empirical risk of the algorithm's output after T iterations. In other words, both the generalization error and the empirical risk provably go down at the same (essentially optimal) rate. Once $T \geq m$, the empirical risk may further decrease to 0, but the generalization error remains at $\tilde{\mathcal{O}}(1/\gamma^2 m)$, which is well-known to be essentially optimal for any learning algorithm in this setting (see for example Hanneke and Kontorovich (2019)).

To prove these results, we also establish more refined, nonasymptotic bounds on the margins attained on the dataset, which are also applicable to other losses. In general, these bounds imply that for any $\alpha \in [0, 1]$, after T iterations, the resulting predictor achieves a

---

2. A bit more precisely, these results imply that for all datasets linearly-separable with margin γ, and for exponentially-tailed losses such as the logistic loss, $\mathbf{w}_T = \mathbf{w}^* \cdot \log(T) + \rho(T)$, where $\rho(T)$ is $\mathcal{O}(1)$ (as a function of T) for datasets in general position, and $\mathcal{O}(\log \log(T))$ always. See for example Gunasekar et al. (2018); Ji and Telgarsky (2019b) for more details.

margin of $\Omega((1-\alpha)\gamma)$ on all but $\tilde{\mathcal{O}}\left(\frac{1}{(\gamma^2 T)^\alpha}\right)$ of the data points. These bounds guarantee that as T increases, for larger and larger portions of the dataset, the predictors achieve a margin of $\Omega(\gamma)$, implying good generalization properties. Furthermore, we prove a lower bound showing that such guarantees are essentially optimal. Finally, although we focus mostly on the exponential and logistic loss, we also discuss the applicability of our results to polynomially-tailed losses in the appendix.

Before continuing, we emphasize that the techniques we use in our upper bounds are not fundamentally new, and similar ideas were employed in previous analyses on the convergence to the max-margin predictor, such as in Ji and Telgarsky (2018b, 2019a) (in fact, some of our results build on these analyses). However, we apply these techniques to a conceptually different question, about the non-asymptotic ability of gradient methods to attain some significant margin. In addition, since we only care about convergence to some large-margin predictor (as opposed to the max-margin predictor), our analysis can be shorter and simpler.

Finally, we note that polynomial-time, non-asymptotic guarantees on the generalization error of unregularized gradient methods were also obtained in Ji and Telgarsky (2019a) and version 2 of Ji and Telgarsky (2018b). However, the former is for nonlinear predictors, and the latter is for one-pass stochastic gradient descent, which is different than the algorithms considered here and where necessarily $T = m$. Moreover, the bounds in both papers have a worse polynomial dependence on the margin γ, compared to our results.

The paper is structured as follows. In the next section, we define some useful notation, and formally describe our setting. In Sec. 3, we provide our positive results about the margin behavior and the generalization error, focusing on gradient flow (for which our analysis is the simplest and completely self-contained). In Sec. 4, we show how similar results can be obtained for gradient descent and stochastic gradient descent. In Sec. 5, focusing for concreteness on gradient descent, we show that our positive result on the margin behavior is essentially tight. In Sec. 6, we discuss how our results can be applied to polynomially-tailed losses, and their implications. A numerical illustration of our results is provided in Sec. 7 Some technical proofs are provided in Appendix A.

2. Preliminaries

We generally let boldfaced letters denote vectors. Given a positive integer n, we let $[n]$ be a shorthand for $\{1, \ldots, n\}$. Given a nonzero vector $\mathbf{w}$, we let

$$\bar{\mathbf{w}} := \mathbf{w}/\|\mathbf{w}\|$$

denote its normalization to unit norm. We use the standard $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ notation to hide constants, and $\tilde{\mathcal{O}}(\cdot), \tilde{\Omega}(\cdot)$ to hide constants and factors polylogarithmic in the problem parameters. $\log(\cdot)$ refers to the natural logarithm, and $\|\cdot\|$ refers to the Euclidean norm.

We consider datasets defined by a set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$, and algorithms which attempt to minimize the empirical risk function, namely

$$\hat{L}(\mathbf{w}) := \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{x}_i^\top \mathbf{w})$$

where $\ell : \mathbb{R} \mapsto \mathbb{R}$ is some loss function[3]. We will utilize the following two assumptions about the dataset and the loss:

Assumption 1 $\max_i \|\mathbf{x}_i\| \leq 1$, *and the dataset is separable with margin* $\gamma \in (0,1]$: *Namely, there exists a unit vector* $\mathbf{w}$ *s.t.* $\min_i \mathbf{x}_i^\top \mathbf{w} \geq \gamma$.

Assumption 2 ℓ *is convex, monotonically decreasing, and has an inverse function[4]* ℓ^{-1} *on the interval* $(0, \ell(0)]$.

Remark 1 *Given a datset, the results in our paper hold for any margin* γ *which satisfy Assumption 1, and in particular the maximum margin (i.e. the largest* γ *for which Assumption 1 hold).*

We note that assumption 1 is without much loss of generality (it simply sets the scaling of the problem). Assumption 2 implies that $\hat{L}$ is convex, and is satisfied for most classification losses. When instantiating our general results, we will focus for concreteness on the logistic loss $\ell(z) = \log(1 + \exp(-z))$ and the exponential loss $\ell(z) = \exp(-z)$. However, our results can be applied to other losses as well.

In our proofs, we will make use of the following well-known facts (see for example Nesterov (2018)): For a convex function f on $\mathbb{R}^d$, it holds for any vectors $\mathbf{u}, \mathbf{v}$ that $f(\mathbf{v}) - f(\mathbf{u}) \leq \nabla f(\mathbf{v})(\mathbf{v} - \mathbf{u})$. Also, if f is a function with μ-Lipschitz gradients, then for any $\mathbf{u}, \mathbf{v}$, $f(\mathbf{v}) \leq f(\mathbf{u}) + \nabla f(\mathbf{u})(\mathbf{v} - \mathbf{u}) + \frac{\mu}{2}\|\mathbf{v} - \mathbf{u}\|^2$.

3. Gradient Flow

In this section, we present positive results on the margin behavior and generalization error of gradient flow. Gradient flow is the standard continuous-time analogue of gradient descent. Although it cannot be implemented precisely in practice, it is a useful idealization of gradient descent, and the method in which it is easiest to present our analysis (the analysis for gradient descent in the next section is just a slight variation). Gradient flow produces a continuous trajectory of vectors $\mathbf{w}(t)$, indexed by a time $t \geq 0$. We define it starting from $\mathbf{0}$, in which case it is specified by the following differential equation:

$$\mathbf{w}(0) = \mathbf{0} \quad , \quad \frac{\partial}{\partial t}\mathbf{w}(t) = -\nabla \hat{L}(\mathbf{w}(t)) . \tag{2}$$

We recall that for linearly separable data, and classification losses which asymptote to zero (such as the exponential and logistic losses), the gradient flow $\mathbf{w}(t)$ generally diverges. However, for classification we are only interested in the direction of the vector (or equivalently, its normalization to a unit vector), and the margin this normalized vector attains on the data.

We now present a general result about the margin behavior of gradient flow, which is applicable to general losses and any vector reached by gradient flow at any time point:

---

3. In the context of binary classification, it is customary to consider labeled data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ and losses of the form $\mathbf{w} \mapsto \ell(y_i \mathbf{x}_i^\top \mathbf{w})$. However, for our purposes we can fold the binary label y_i inside $\mathbf{x}_i$, and treat this as a single vector.
4. Namely, for any $z \in (0, \infty)$, there is a unique $p = \ell^{-1}(z)$ such that $\ell(p) = z$.

Theorem 2 *Under assumptions 1 and 2, let* $\mathbf{w} \neq \mathbf{0}$ *be some point reached by gradient flow (as specified in Eq. (2)), such that* $\hat{L}(\mathbf{w}) = \epsilon$ *for some* $\epsilon \in (0, \ell(0)]$. *Then for any* $p \in \left[\frac{\epsilon}{\ell(0)}, 1\right]$, *for at least* $(1-p)m$ *of the indices* $i \in [m]$,

$$\mathbf{x}_i^\top \bar{\mathbf{w}} \; > \; \frac{\gamma}{2} \cdot \frac{\ell^{-1}(\epsilon/p)}{\ell^{-1}(\epsilon)} \; .$$

Intuitively, we expect the empirical risk $\hat{L}(\mathbf{w}(t))$ to decay with t (as we instantiate for specific losses later on). Computing the corresponding bounds on ϵ and plugging into the above, we can get guarantees on how many points in our dataset achieve a certain margin. Note that since ℓ^{-1} is monotonically decreasing, the margin lower bound in the theorem is always at most $\gamma/2$. Since under assumption 1 the maximal margin is at least γ, this result cannot be used to recover the asymptotic convergence to the max-margin predictor, shown by previous results. However, as discussed in the introduction, this is not our focus here: We ask about the time to converge to some large-margin predictor which generalizes well, not necessarily the max-margin predictor. For that purpose, as we will see later, a margin lower bound of $\Omega(\gamma)$ is perfectly adequate.

To prove Thm. 2, we will need the following key lemma, which bounds the norm of the points along the trajectory of gradient flow, in terms of the value of $\hat{L}$. The proof is short and relies only on the convexity of $\hat{L}$:

Lemma 3 *Fix some* $T \geq 0$, *and let* $\mathbf{w}^*$ *be any vector such that* $\hat{L}(\mathbf{w}^*) \leq \hat{L}(\mathbf{w}(T))$. *Then* $\sup_{t \in [0,T]} \|\mathbf{w}(t)\| \leq 2\|\mathbf{w}^*\|$.

Proof By definition of gradient flow and the chain rule, we have $\frac{\partial}{\partial t}\hat{L}(\mathbf{w}(t)) = -\|\nabla\hat{L}(\mathbf{w}(t))\|^2 \leq 0$, so $\hat{L}(\mathbf{w}(t))$ is monotonically decreasing in t. As a result, for any t, $\hat{L}(\mathbf{w}(t)) \geq \hat{L}(\mathbf{w}(T)) \geq \hat{L}(\mathbf{w}^*)$. By convexity of $\hat{L}$, is follows that

$$\nabla\hat{L}(\mathbf{w}(t))^\top(\mathbf{w}(t) - \mathbf{w}^*) \; \geq \; \hat{L}(\mathbf{w}(t)) - \hat{L}(\mathbf{w}^*) \geq 0 \; .$$

Using this inequality, the definition of gradient flow and the chain rule, it follows that

$$\frac{\partial}{\partial t}\|\mathbf{w}(t) - \mathbf{w}^*\|^2 \; = \; -2\nabla\hat{L}(\mathbf{w}(t))^\top(\mathbf{w}(t) - \mathbf{w}^*) \leq 0 \; .$$

Therefore, $\|\mathbf{w}(t) - \mathbf{w}^*\|$ is monotonically decreasing in $t \in [0, T]$, so it is at most $\|\mathbf{w}(0) - \mathbf{w}^*\| = \|\mathbf{w}^*\|$. Thus, by the triangle inequality, $\|\mathbf{w}(t)\| \leq \|\mathbf{w}(t) - \mathbf{w}^*\| + \|\mathbf{w}^*\| \leq 2\|\mathbf{w}^*\|$. ∎

Proof [Proof of Thm. 2] Let $\mathbf{w}_0, \|\mathbf{w}_0\| = 1$ be a max-margin separator, so that $\min_i \mathbf{x}_i^\top \mathbf{w}_0 \geq \gamma$. Define $\mathbf{w}^* := \frac{\ell^{-1}(\epsilon)}{\gamma}\mathbf{w}_0$, which has norm $\ell^{-1}(\epsilon)/\gamma$, and note that since $\ell^{-1}(\epsilon)$ is non-negative,

$$\max_i \ell(\mathbf{x}_i^\top \mathbf{w}^*) \; = \; \ell\left(\min_i \frac{\ell^{-1}(\epsilon)}{\gamma}\mathbf{x}_i^\top \mathbf{w}_0\right) \; \leq \; \ell(\ell^{-1}(\epsilon) \cdot 1) \; = \; \epsilon \; .$$

This implies $\hat{L}(\mathbf{w}^*) = \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{x}_i^\top \mathbf{w}^*) \leq \epsilon$. Combined with Lemma 3, we get that

$$\|\mathbf{w}\| \; \leq \; 2\|\mathbf{w}^*\| \; = \; \frac{2\ell^{-1}(\epsilon)}{\gamma} \; . \tag{3}$$

Since $\hat{L}(\mathbf{w}) = \epsilon$, we get that $\mathbb{E}_i[\ell(\mathbf{x}_i^\top \mathbf{w})] \le \epsilon$, where i is uniformly distributed on $[m]$. By Markov's inequality and Eq. (3), it follows that

$$
\begin{aligned}
p \ &\ge\ \Pr_i\left(\ell(\mathbf{x}_i^\top \mathbf{w}) \ge \frac{\epsilon}{p}\right) \ =\ \Pr_i\left(\mathbf{x}_i^\top \mathbf{w} \le \ell^{-1}(\epsilon/p)\right) \ =\ \Pr_i\left(\mathbf{x}_i^\top \bar{\mathbf{w}} \le \frac{\ell^{-1}(\epsilon/p)}{\|\mathbf{w}\|}\right) \\
&\ge\ \Pr_i\left(\mathbf{x}_i^\top \bar{\mathbf{w}} \le \frac{\gamma \ell^{-1}(\epsilon/p)}{2\ell^{-1}(\epsilon)}\right) \ .
\end{aligned}
$$

$\blacksquare$

For concreteness, let us now apply Thm. 2 to the case of the exponential loss, $\ell(z) = \exp(-z)$. In order to get an interesting guarantee, we will utilize the following simple non-asymptotic guarantee on the decay of $\hat{L}(\mathbf{w}(T))$ as a function of T:

Lemma 4 *Under assumption 1, if ℓ is the exponential loss, then $\hat{L}(\mathbf{w}(T)) \le \frac{1}{\gamma^2 T}$ for any $T > 0$.*

Proof Let $\mathbf{w}_0$ be a max-margin unit vector, so that $\min_i \mathbf{x}_i^\top \mathbf{w}_0 \ge \gamma$. By definition of gradient flow, the chain rule and Cauchy-Schwartz,

$$
\begin{aligned}
\frac{\partial}{\partial t} \hat{L}(\mathbf{w}(t)) \ &=\ -\|\nabla \hat{L}(\mathbf{w}(t))\|^2 \ \le\ -\left(\nabla \hat{L}(\mathbf{w}(t))^\top \mathbf{w}_0\right)^2 \ =\ -\left(\frac{1}{m}\sum_{i=1}^m \ell'(\mathbf{x}_i^\top \mathbf{w}(t))\mathbf{x}_i^\top \mathbf{w}_0\right)^2 \\
&\le\ -\left(\frac{1}{m}\sum_{i=1}^m \ell'(\mathbf{x}_i^\top \mathbf{w}(t))\gamma\right)^2 \ \overset{(*)}{=}\ -\left(\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{x}_i^\top \mathbf{w}(t))\gamma\right)^2 \ =\ -\gamma^2 \hat{L}(\mathbf{w}(t))^2 \ ,
\end{aligned}
$$

where in $(*)$ we used the fact that for the exponential loss, $\ell(z) = -\ell'(z)$ for all z. Consider now the function $f(t) := \gamma^2 t \cdot \hat{L}(\mathbf{w}(t))$. We clearly have $f(0) = 0$, and by differentiation and the inequality above, it is easily verified that $f'(t) \le 0$ whenever $f(t) \ge 1$. Since $f(t)$ is continuous, we must have $f(t) \le 1$ for all t, hence $\hat{L}(\mathbf{w}(t)) \le \frac{1}{\gamma^2 t}$. $\blacksquare$

Using this lemma, we get the following corollary of Thm. 2 for the exponential loss:

Corollary 5 *Under assumption 1, if ℓ is the exponential loss, then for any $T > 1/\gamma^2$ and any $\alpha \in [0,1]$, it holds that $\mathbf{x}_i^\top \bar{\mathbf{w}}(T) > \frac{1-\alpha}{2} \cdot \gamma$ for at least $(1 - (\gamma^2 T)^{-\alpha})m$ of the indices $i \in [m]$.*

Proof By Lemma 4, we know that at time T, $\hat{L}(\mathbf{w}(T)) = \epsilon$ for some $\epsilon \le 1/\gamma^2 T < 1 = \ell(0)$ (which implies $\mathbf{w}(T) \ne \mathbf{0}$). Applying Thm. 2, and noting that $\ell^{-1}(z) = \log(1/z)$ and $\ell(0) = 1$, we get that for any $p \in [1/\gamma^2 T, 1]$, for at least $(1-p)m$ of the indices i,

$$
\mathbf{x}_i^\top \bar{\mathbf{w}}_T \ >\ \frac{\gamma}{2} \cdot \frac{\log(p/\epsilon)}{\log(1/\epsilon)} \ =\ \frac{\gamma}{2} \cdot \left(1 - \frac{\log(1/p)}{\log(1/\epsilon)}\right) \ \ge\ \frac{\gamma}{2} \cdot \left(1 - \frac{\log(1/p)}{\log(\gamma^2 T)}\right) \ .
$$

In particular, picking $p = (\gamma^2 T)^{-\alpha}$ for some $\alpha \in [0,1]$, the result follows. $\blacksquare$

6

Note that if $T > m^{1/\alpha}/\gamma^2$, then $(\gamma^2 T)^{-\alpha} < \frac{1}{m}$, which implies that for *all* $i \in [m]$, $\mathbf{x}_i^\top \bar{\mathbf{w}}(T)$ is at least $\frac{1-\alpha}{2} \cdot \gamma$. However, even for smaller values of T, the theorem provides guarantees on the margin attained on most points in the dataset. Moreover, using standard margin-based generalization bounds for binary classification, this theorem implies that the predictors returned by gradient flow achieve low generalization error, uniformly over all sufficiently large T:

Theorem 6 *Let $\mathcal{D}$ be some distribution over $\{\mathbf{x} : \|\mathbf{x}\| \leq 1\} \times \{-1, +1\}$, such that there exists a unit vector $\mathbf{w}_0$ and $\gamma > 0$ satisfying $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}(y\mathbf{x}^\top \mathbf{w}_0 \geq \gamma) = 1$. If we sample m points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ i.i.d. from $\mathcal{D}$, and run gradient flow on $\mathbf{w} \mapsto \frac{1}{m}\sum_{i=1}^m \ell(y_i\mathbf{x}_i^\top \mathbf{w})$, where ℓ is the exponential loss, then for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the sample, it holds for any time T that*

$$\Pr_{(\mathbf{x},y)\sim D}(\text{sign}(\mathbf{x}^\top \mathbf{w}(T)) \neq y) \leq \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 T} + \frac{1}{\gamma^2 m}\right) ,$$

where the $\tilde{\mathcal{O}}$ notation hides universal constants and factors polylogarithmic in $\gamma^2 m$ and $1/\delta$.

As discussed in the introduction, this is essentially the best behavior we can hope for (up to log factors): For $T \leq m$, the generalization error decays as $\tilde{\mathcal{O}}(1/\gamma^2 T)$, which is also the bound on the empirical risk (see Lemma 4). Once $T \geq m$, the generalization error becomes $\tilde{\mathcal{O}}(1/\gamma^2 m)$, and stays there regardless of how large T is.

Proof [Proof of Thm. 6] The bound in the theorem is vacuous when $\gamma^2 m \leq 1$ or $\gamma^2 T \leq 1$, so we will assume without loss of generality that both quantities are larger than 1.

Standard margin-based generalization bounds (e.g. McAllester (2003)) imply that in our setting, if we pick m points i.i.d., then with probability at least $1 - \delta$, any vector $\mathbf{w}$ for which $\max_i y_i\mathbf{x}_i^\top \bar{\mathbf{w}} \geq \hat{\gamma}$ for all but pm of the points satisfies

$$\Pr_{(\mathbf{x},y)\sim D}(\text{sign}(\mathbf{w}^\top \mathbf{x}) \neq y) \leq p + \tilde{\mathcal{O}}\left(\sqrt{p \cdot \frac{1}{\hat{\gamma}^2 m}} + \frac{1}{\hat{\gamma}^2 m}\right) \leq \tilde{\mathcal{O}}\left(p + \frac{1}{\hat{\gamma}^2 m}\right) ,$$

where the $\tilde{\mathcal{O}}$ hides universal constants and factors polylogarithmic in $1/\delta$ and $\gamma^2 m$. In particular, this can be applied uniformly for $\mathbf{w}(T)$ for any T. By Corollary 5, we can substitute $p = (\gamma^2 T)^{-\alpha}$ and $\hat{\gamma} = (1 - \alpha)\gamma/2$, to get that with probability at least $1 - \delta$,

$$\Pr_{(\mathbf{x},y)\sim D}(\text{sign}(\mathbf{x}^\top \mathbf{w}(T)) \neq y) \leq \tilde{\mathcal{O}}\left(\frac{1}{(\gamma^2 T)^\alpha} + \frac{1}{(1-\alpha)^2\gamma^2 m}\right) . \tag{4}$$

This holds for any $\alpha \in [0, 1]$. In particular, if $\gamma^2 T \geq (\gamma^2 m)^2$, pick $\alpha = 1/2$, in which case Eq. (4) is at most

$$\tilde{\mathcal{O}}\left(\frac{1}{(\gamma^2 T)^{1/2}} + \frac{1}{\gamma^2 m}\right) = \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 m}\right) ,$$

and if $\gamma^2 T < (\gamma^2 m)^2$, pick $\alpha = 1 - \frac{1}{2\log(\gamma^2 m)}$, in which case Eq. (4) is at most

$$\tilde{\mathcal{O}}\left(\frac{(\gamma^2 T)^{1/2\log(\gamma^2 m)}}{\gamma^2 T} + \frac{1}{\gamma^2 m}\right) \leq \tilde{\mathcal{O}}\left(\frac{(\gamma^2 m)^{1/\log(\gamma^2 m)}}{\gamma^2 T} + \frac{1}{\gamma^2 m}\right) = \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 T} + \frac{1}{\gamma^2 m}\right) ,$$

where we used the fact that $z^{1/\log(z)} = \exp(\log(z)/\log(z)) = \exp(1)$. Combining the two cases, the result follows. ■

4. Gradient Descent and Stochastic Gradient Descent

Having discussed gradient flow, we show in this section how essentially identical results can be obtained for gradient descent and stochastic gradient descent.

4.1 Gradient Descent

Gradient descent, which is probably the simplest and most well-known gradient method, optimizes $\hat{L}$ by initializing at some point $\mathbf{w}_1$ (which we will take to be $\mathbf{0}$), and performing iterations of the form $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \hat{L}(\mathbf{w}_t)$, where $\eta_1, \eta_2, \ldots$ are step size parameters. We will utilize the following standard assumption:

Assumption 3 *The derivative of ℓ is μ-Lipschitz, and $0 < \eta_t \leq 1/\mu$ for all t.*

Inspecting the analysis for gradient flow from the previous section, we note that we relied on the algorithm's structure only at two points: In Lemma 3, to bound the norm of the points along the trajectory, and in Lemma 4, to upper bound the values of $\hat{L}$. Fortunately, we can provide analogues of these two lemmas for gradient descent:

Lemma 7 *Under assumptions 1,2 and 3, fix some index $T \geq 1$, and let $\mathbf{w}^*$ be any vector such that $\hat{L}(\mathbf{w}^*) \leq \hat{L}(\mathbf{w}_T)$. Then $\max_{t \in [T]} \|\mathbf{w}_t\| \leq 2\|\mathbf{w}^*\|$.*

Lemma 8 *(Ji and Telgarsky (2018b, Theorem 3.1)) If ℓ is the logistic loss, then under assumption 3, gradient descent with step size $\eta = 1$ satisfies $\hat{L}(\mathbf{w}_T) \leq \frac{1}{T} + \frac{\log^2(T)}{2\gamma^2 T}$ for any T.*

The proof of Lemma 7 (which is a slight variation on the proof of Lemma 3) appears in Appendix A.

With these lemmas, we can prove analogues of the theorems from the previous section, this time for gradient descent and for the logistic loss. The resulting bounds are identical up to constants and logarithmic factors:

Theorem 9 *Under assumptions 1, 2 and 3, let $\mathbf{w} \neq \mathbf{0}$ be some point reached by gradient descent, such that $\hat{L}(\mathbf{w}) = \epsilon$ for some $\epsilon \in (0, \ell(0)]$. Then for any $p \in \left[\frac{\epsilon}{\ell(0)}, 1\right]$, for at least $(1-p)m$ of the indices $i \in [m]$,*

$$\mathbf{x}_i^\top \bar{\mathbf{w}} > \frac{\gamma}{2} \cdot \frac{\ell^{-1}(\epsilon/p)}{\ell^{-1}(\epsilon)} \ .$$

Corollary 10 *Under assumption 1, if ℓ is the logistic loss, and we use a fixed step size of $\eta_t = 1$ for all t, then for any $T > 4$ such that $\frac{\log^2(T)}{\gamma^2 T} < \ell(0)$, and any $\alpha \in [0,1]$, the gradient descent iterates satisfy $\mathbf{x}_i^\top \bar{\mathbf{w}}_T > \frac{1-\alpha}{2} \cdot \gamma$ for at least $\left(1 - 2\left(\frac{\log^2(T)}{\gamma^2 T}\right)^\alpha\right)m$ of the indices $i \in [m]$.*

Theorem 11 *Let $\mathcal{D}$ be some distribution over $\{\mathbf{x} : \|\mathbf{x}\| \leq 1\} \times \{-1, +1\}$, such that there exists a unit vector $\mathbf{w}_0$ and $\gamma > 0$ satisfying $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}(y\mathbf{x}^\top\mathbf{w}_0 \geq \gamma) = 1$. If we sample m points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ i.i.d. from $\mathcal{D}$, and run gradient descent with fixed step sizes $\eta_t = 1$ on $\mathbf{w} \mapsto \frac{1}{m}\sum_{i=1}^m \ell(y_i\mathbf{x}_i^\top\mathbf{w})$, where ℓ is the logistic loss, then for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the sample, it holds for any iteration T that*

$$\Pr_{(\mathbf{x},y)\sim D}(\mathrm{sign}(\mathbf{x}^\top\mathbf{w}_T) \neq y) \;\leq\; \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 T} + \frac{1}{\gamma^2 m}\right) \;,$$

where the $\tilde{\mathcal{O}}$ notation hides universal constants and factors polylogarithmic in $\gamma^2 m$, T and $1/\delta$.

The results can be easily generalized to other step size strategies. The proofs are essentially identical to the proofs from the previous section, except that we use Lemma 7 and 8 instead of Lemmas 3 and 4. In particular, the proof of Thm. 9 is identical to the proof of Thm. 2; the proof of Corollary 10 is nearly identical to the proof of Corollary 5 (and is provided in Appendix A for completeness); and the proof of Thm. 11 is identical to the proof of Thm. 6, except that we plug in Corollary 10 instead of Corollary 5, and thus have some additional $\log(T)$ factors which get absorbed into the $\tilde{\mathcal{O}}()$ notation.

4.2 Stochastic Gradient Descent

We now turn to discuss the stochastic gradient descent (SGD) algorithm, perhaps the main workhorse of modern machine learning methods. We consider the simplest version of SGD for minimizing $\hat{L}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{x}_i^\top\mathbf{w})$: We initialize $\mathbf{w}_1$ at the origin $\mathbf{0}$, and for any $t \geq 1$, define $\mathbf{w}_{t+1} := \mathbf{w}_t - \eta\ell'(\mathbf{x}_{i_t}^\top\mathbf{w}_t)\mathbf{x}_{i_t}$, where $i_t \in [m]$ is chosen independently and uniformly at random (so that in expectation, $\mathbb{E}_{i_t}[\mathbf{w}_{t+1}] = \mathbf{w}_t - \eta\nabla\hat{L}(\mathbf{w}_t)$, similar to the gradient descent update). We assume that at the end of T iterations, the algorithm returns the average of the iterates obtained so far, $\mathbf{v}_T := \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$.

To avoid yet another (and more complicated) repetition of the analysis from the previous section, we will take a somewhat different route, focusing on the logistic loss and fixed step sizes $\eta_t = 1$, which allows us to directly utilize some existing results in the literature to get a bound on the margin behavior of SGD (analogous to Corollary 5 for gradient flow and Corollary 10 for gradient descent). We note that the analysis can be easily generalized to other constant step sizes.

Theorem 12 *Under assumption 1, if ℓ is the logistic loss, and we use step sizes $\eta_t = 1$, then for any $T \geq 2/\gamma^2$ and any $\alpha \in [0, 1]$, the SGD iterates satisfy $\mathbf{x}_i^\top\bar{\mathbf{v}}_T > \frac{1-\alpha}{5}\cdot\gamma$ for at least $(1-\delta)m$ of the indices $i \in [m]$, where δ is a nonnegative random variable (dependent on the randomness of SGD), whose expectation is at most $\frac{8+4\log^2(\gamma^2 T)}{3\gamma^2 T^\alpha}$.*

Proof We will utilize the following easily-verified facts about the logistic loss: It is nonnegative, its gradient is $\frac{1}{4}$-Lipschitz, and its inverse is $\ell^{-1}(z) := \log(1/(\exp(z) - 1))$, which is between $\log(1/z)$ and $\log(1/2z)$ for all $z \in [0, 1]$.

Let $\mathbf{w}_0, \|\mathbf{w}_0\| = 1$ be a max-margin separator, so that $\min_i \mathbf{x}_i^\top\mathbf{w}_0 \geq \gamma$, and define (similarly to the proof of Thm. 6) $\mathbf{w}^* := \frac{\ell^{-1}(\epsilon)}{\gamma}\mathbf{w}_0$, where $\epsilon \in (0, \frac{1}{2}]$ will be chosen later

(note that $\ell^{-1}(\epsilon) > 0$ in this regime). It is easily verified that $\max_i \ell(\mathbf{x}_i^\top \mathbf{w}^*) \leq \epsilon$, and therefore $\hat{L}(\mathbf{w}^*) \leq \epsilon$. Since ℓ is non-negative and with a $\frac{1}{4}$-Lipschitz derivative, we can use Theorem 14.13 from Shalev-Shwartz and Ben-David (2014) on the convergence of SGD for such losses to get

$$\mathbb{E}[\hat{L}(\mathbf{v}_T)] \leq \frac{1}{1 - \frac{1}{4}}\left(\hat{L}(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|^2}{2T}\right) \leq \frac{4}{3}\left(\epsilon + \frac{(\ell^{-1}(\epsilon))^2}{2\gamma^2 T}\right).$$

In particular, picking $\epsilon = \frac{1}{\gamma^2 T}$ (which is in $(0, \frac{1}{2}]$ by our assumption $T \geq 2/\gamma^2$), we get

$$\mathbb{E}[\hat{L}(\mathbf{v}_T)] \leq \frac{4}{3\gamma^2 T}\left(1 + \frac{(\ell^{-1}(1/\gamma^2 T))^2}{2}\right) \leq \frac{4}{3\gamma^2 T}\left(1 + \frac{1}{2}\log^2(\gamma^2 T)\right).$$

To simplify notation, define $\Delta(T, \gamma)$ to be the expression in the right-hand side above. We also note that the left-hand side equals $\mathbb{E}_{\mathbf{v}_T, i}[\ell(\mathbf{x}_i^\top \mathbf{v}_T)]$, where i is uniformly distributed in $[m]$. Thus, by Markov's inequality, for any $p > 0$, we have

$$p \geq \Pr_{i, \mathbf{v}_T}\left(\ell(\mathbf{x}_i^\top \mathbf{v}_T) \geq \frac{\Delta(T, \gamma)}{p}\right) = \Pr_{i, \mathbf{v}_T}\left(\mathbf{x}_i^\top \mathbf{v}_T \leq \ell^{-1}\left(\frac{\Delta(T, \gamma)}{p}\right)\right).$$

According to Theorem 2.1 in version 2 of Ji and Telgarsky (2018b), the iterates of SGD on the logistic loss satisfy deterministically $\max_{t \in [T]} \|\mathbf{w}_t\| \leq \frac{2\log(T)}{\gamma} + 2$, so by Jensen's inequality, $\|\mathbf{v}_T\| \leq \frac{1}{T}\sum_{t=1}^T \|\mathbf{w}_t\| \leq \frac{2\log(T)}{\gamma} + 2$. Combining this with the displayed equation above, we get that

$$p \geq \Pr_{i, \mathbf{v}_T}\left(\mathbf{x}_i^\top \bar{\mathbf{v}}_T \leq \frac{1}{2 + 2\log(T)/\gamma} \cdot \ell^{-1}\left(\frac{\Delta(T, \gamma)}{p}\right)\right).$$

Choosing $p = 2T^{1-\alpha}\Delta(T, \gamma) = \frac{8 + 4\log^2(\gamma^2 T)}{3\gamma^2 T^\alpha}$ for some $\alpha \in [0, 1]$, and substituting into the above, we get that

$$\frac{8 + 4\log^2(\gamma^2 T)}{3\gamma^2 T^\alpha} \geq \Pr_{i, \mathbf{v}_T}\left(\mathbf{x}_i^\top \bar{\mathbf{v}}_T \leq \frac{1}{2 + 2\log(T)/\gamma} \cdot \ell^{-1}\left(\frac{T^{\alpha-1}}{2}\right)\right)$$

$$\geq \Pr_{i, \mathbf{v}_T}\left(\mathbf{x}_i^\top \bar{\mathbf{v}}_T \leq \frac{1}{2 + 2\log(T)/\gamma} \cdot \log\left(T^{1-\alpha}\right)\right)$$

$$= \Pr_{i, \mathbf{v}_T}\left(\mathbf{x}_i^\top \bar{\mathbf{v}}_T \leq \frac{(1-\alpha)\gamma}{2} \cdot \frac{\log(T)}{\gamma + \log(T)}\right).$$

Since $T > \frac{1}{\gamma^2} \geq 1$, we have $\frac{\log(T)}{\gamma + \log(T)} \geq \frac{\log(T)}{1 + \log(T)} \geq \frac{\log(2)}{1 + \log(2)} > \frac{2}{5}$. Plugging into the above, we get

$$\frac{8 + 4\log^2(\gamma^2 T)}{3\gamma^2 T^\alpha} \geq \Pr_{i, \mathbf{v}_T}\left(\mathbf{x}_i^\top \bar{\mathbf{v}}_T \leq \frac{(1-\alpha)\gamma}{5}\right).$$

To complete the proof, we note that if we let A_i denote the event that $\mathbf{x}_i^\top \bar{\mathbf{v}}_T \leq \frac{(1-\alpha)\gamma}{5}$, and $\mathbf{1}_{A_i}$ the indicator function of the event A_i, then the above implies

$$\frac{8 + 4\log^2(\gamma^2 T)}{3\gamma^2 T^\alpha} \geq \mathbb{E}_{i, \mathbf{v}_T}[\mathbf{1}_{A_i}] = \mathbb{E}_{\mathbf{v}_T}\mathbb{E}_i[\mathbf{1}_{A_i}] = \mathbb{E}_{\mathbf{v}_T}\left[\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{A_i}\right].$$

Thus, letting $\delta = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{A_i}$, the theorem follows. ■

Using this theorem, we get the following generalization error bound for SGD, which is a direct analogue of the error bounds we obtained for gradient flow and gradient descent:

Theorem 13 *Let $\mathcal{D}$ be some distribution over $\{\mathbf{x} : \|\mathbf{x}\| \leq 1\} \times \{-1, +1\}$, such that there exists a unit vector $\mathbf{w}_0$ and $\gamma > 0$ satisfying $\Pr_{(\mathbf{x},y) \sim \mathcal{D}}(y\mathbf{x}^\top \mathbf{w}_0 \geq \gamma) = 1$. Suppose we sample m points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ i.i.d. from $\mathcal{D}$, and run SGD (with step sizes $\eta_t = 1$) on $\mathbf{w} \mapsto \frac{1}{m} \sum_{i=1}^{m} \ell(y_i \mathbf{x}_i^\top \mathbf{w})$, where ℓ is the logistic loss. Then for any T,*

$$\mathbb{E}\left[\Pr_{(\mathbf{x},y) \sim D}(\text{sign}(\mathbf{x}^\top \mathbf{v}_T) \neq y) \right] \leq \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 T} + \frac{1}{\gamma^2 m} \right) ,$$

where the $\tilde{\mathcal{O}}$ notation hides universal constants and factors polylogarithmic in $T, m, 1/\gamma$, and the expectation is over the randomness of the SGD algorithm.

Proof Using identical arguments as in the proof of Thm. 6 (except using Thm. 12 instead of Corollary 5, and the fact that high-probability bounds imply a bound on the expectation), we get that

$$\mathbb{E}\left[\Pr_{(\mathbf{x},y) \sim D}(\text{sign}(\mathbf{x}^\top \mathbf{w}(T)) \neq y) \right] \leq \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 T^\alpha} + \frac{1}{(1-\alpha)^2 \gamma^2 m} \right) . \tag{5}$$

This holds for any $\alpha \in [0, 1]$. In particular, if $T \geq m^2$, pick $\alpha = 1/2$, in which case Eq. (5) is at most

$$\tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 T^{1/2}} + \frac{1}{\gamma^2 m} \right) = \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 m} \right) ,$$

and if $T < m^2$, pick $\alpha = 1 - \frac{1}{\log(m^2)}$, in which case Eq. (5) is at most

$$\tilde{\mathcal{O}}\left(\frac{T^{1/\log(m^2)}}{\gamma^2 T} + \frac{1}{\gamma^2 m} \right) \leq \tilde{\mathcal{O}}\left(\frac{(m^2)^{1/\log(m^2)}}{\gamma^2 T} + \frac{1}{\gamma^2 m} \right) = \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 T} + \frac{1}{\gamma^2 m} \right) .$$

Combining the two cases, the result follows. ■

Finally, we remark that Thm. 13 only bounds the expectation of the error probability of $\mathbf{v}_T$. It is likely that using more sophisticated concentration tools, one can obtain a high-probability bound. However, this would require a more involved analysis, and is left for future work.

5. Tightness

In the previous sections, we provided bounds on the margin behavior of gradient flow, gradient descent and SGD, which all have the following form (ignoring constants and log factors): After T iterations, we get a predictor which achieves $\Omega(\gamma)$ margin on at least $(1 - (\gamma^2 T)^{-\alpha})m$ of the m data points, where $\alpha \in (0, 1)$ is a constant arbitrarily close to one. It is natural to ask whether this result can be improved.

In this section, we show that this result is essentially tight: After T iterations, it is impossible to guarantee any positive margin on more than $(1-p)m$ of the points, when p is much less than $(\gamma^2 T)^{-1}$. For concreteness, we prove this for gradient descent and the logistic loss, although the analysis can be extended to other gradient methods and losses:

Theorem 14 *For any positive integers m, T and any $\gamma \in \left(0, \frac{1}{8}\right]$, there exists a dataset of m points in $\mathbb{R}^2$ satisfying assumption 1, such that gradient descent using the logistic loss and any step size $\eta \leq 1$ must satisfy $\mathbf{x}_i^\top \mathbf{w}_T \leq 0$ for at least $\lfloor \frac{m}{26 \max\{1, \gamma^2 T\}} \rfloor$ of the data points.*

Since $\mathbf{x}_i^\top \mathbf{w}_T \leq 0$ translates to $\ell(\mathbf{x}_i^\top \mathbf{w}_T) \geq \Omega(1)$, the theorem also implies that the $\tilde{\mathcal{O}}(1/\gamma^2 T)$ bounds on the empirical risk shown earlier are tight up to logarithmic factors. It also implies that the percentage of misclassified points ($\mathbf{x}_i^\top \mathbf{w}_T \leq 0$) cannot decrease at a better rate. In addition, the lower bound implies that if we want to get any positive margin on all m data points, the number of iterations T must be at least $\Omega(m/\gamma^2)$ in the worst case.

A lower bound related to ours appears in Ji and Telgarsky (2019b), where the authors show that if $\mathbf{w}^*$ is the max-margin unit-norm predictor, then $\|\bar{\mathbf{w}}_T - \mathbf{w}^*\| \geq \Omega(\log(m)/\log(T))$. This translates to a $T \geq \Omega(m)$ requirement to get a non-trivial guarantee on the direction of $\bar{\mathbf{w}}_T$. However, this is a somewhat different objective than ours, and moreover, their lower bound does not specify the dependence on the margin parameter γ.

Proof [Proof of Thm. 14]

We will assume without loss of generality that $T \geq \frac{1}{\gamma^2}$ (or equivalently, $\gamma^2 T \geq 1$), in which case the lower bound we need to prove is $\lfloor \frac{m}{26\gamma^2 T} \rfloor$. Otherwise, if $T < \frac{1}{\gamma^2}$, we can simply apply the construction below with the larger margin parameter $\hat{\gamma} := \frac{1}{\sqrt{T}}$, which by definition satisfies $T \geq \frac{1}{\hat{\gamma}^2}$, and uses a dataset separable with margin $\hat{\gamma} \geq \gamma$, so assumption 1 still holds with margin parameter γ.

We will also assume without loss of generality that $\lfloor \frac{m}{26\gamma^2 T} \rfloor > 0$ (otherwise the theorem trivially holds), and fix $\epsilon := \frac{1}{m} \lfloor \frac{m}{26\gamma^2 T} \rfloor$ (which is positive and less than $\frac{1}{26\gamma^2 T} \leq \frac{1}{26}$).

Consider a dataset consisting of a majority group of $(1-\epsilon)m$ points of the form $(1, 0)$, and a minority group of ϵm points of the form $\left(-\frac{1}{2}, 3\gamma\right)$ (see Figure 1 for a sketch of the construction and proof idea). It is easily verified that all these points are contained in the unit ball, and that the vector $\mathbf{w} = \left(\gamma, \frac{1}{2}\right)$ satisfies $\mathbf{x}^\top \bar{\mathbf{w}} \geq \mathbf{x}^\top \mathbf{w} \geq \gamma$ for any $\mathbf{x}$ in the dataset. Thus, assumption 1 holds. Moreover, the empirical risk function equals

$$\hat{L}(\mathbf{w}) = \hat{L}(w(1), w(2)) = (1-\epsilon) \cdot \ell(w(1)) + \epsilon \cdot \ell\left(-\frac{1}{2} \cdot w(1) + 3\gamma \cdot w(2)\right),$$

where $\ell(z) := \log(1 + \exp(-z))$ is the logistic loss.

Suppose by contradiction that $\mathbf{x}^\top \mathbf{w}_T \leq 0$ for less than ϵm of the points $\mathbf{x}$ in the dataset. Since there are ϵm points of the form $\mathbf{x} = \left(-\frac{1}{2}, 3\gamma\right)$, this implies that all these points are positively correlated with $\mathbf{w}_T$, that is $\mathbf{x}^\top \mathbf{w}_T = -\frac{1}{2} \cdot w_T(1) + 3\gamma \cdot w_T(2) > 0$. The lemma below implies that this cannot happen unless $T \geq \frac{19}{480\gamma^2 \epsilon}$. But since $\frac{19}{480\gamma^2 \epsilon} \geq \frac{19}{480\gamma^2(1/26\gamma^2 T)} > T$, we get $T > T$, a contradiction.
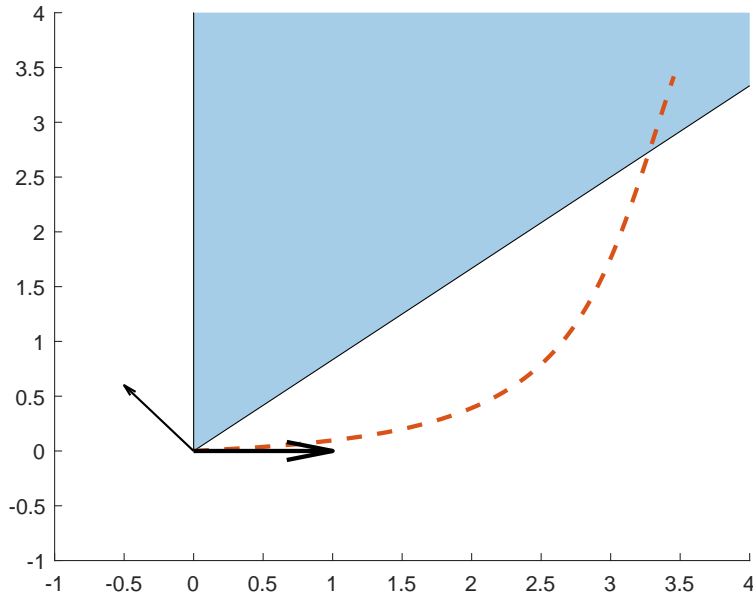
Figure 1: Illustration of the proof of Thm. 14, for $\gamma = \frac{1}{5}$. The dataset consists of 90% points at $(1,0)$ (thick black arrow) and 10% points at $(-1/2, 3\gamma)$ (thin black arrow). The gradient descent trajectory (with $\eta = 1$, starting from the origin) is the dotted red line, and the shaded blue region are the vectors which achieves a positive margin on more than 90% of the data points (or equivalently for this construction, get a positive margin on all data points). Initially, the gradient descent trajectory is mostly influenced by the points at $(1,0)$, and only once a sufficiently large margin is achieved on them, the influence of the few points at $(-1/2, 3\gamma)$ begins to manifest, and the trajectory curves towards the blue region. Best viewed in color.

Lemma 15 *Fix some ϵ and γ in $\left(0, \frac{1}{8}\right]$. Consider gradient descent on the bivariate function*

$$\hat{L}(r, s) := (1 - \epsilon) \cdot \ell(r) + \epsilon \cdot \ell\left(-\frac{1}{2}r + 3\gamma s\right) ,$$

where ℓ is the logistic loss, using any step size $\eta \leq 1$, starting from $r_1 = s_1 = 0$ and producing a sequence of iterates $(r_2, s_2), (r_3, s_3), \ldots$. Then if $-\frac{1}{2}r_t + 3\gamma s_t > 0$, we must have $t \geq \frac{19}{480\gamma^2\epsilon}$.

The proof of the lemma is based on a technical calculation, and appears in Appendix A.

∎

13

6. Polynomially-Tailed Losses

In our paper, we focused mostly on the exponential and logistic losses, which both have exponentially decaying tails. Nacson et al. (2019a) showed that such a tail is important to get asymptotic convergence to a max-margin solution: If $\ell(z)$ decays polynomially with z, then the convergence may not be to a max-margin solution.

Despite this, in the spirit of our previous results, one may still conjecture that we converge to some large-margin solution (though not a max-margin one). Recently, Dudik et al. (2020, Remark 8) showed this holds, albeit in a weak sense: If the loss $\ell(z)$ decays as z^{-b} for some $b > 0$, then the asymptotic margin over the dataset is $\Omega(m^{-1/(b+1)})$, and this is tight in general.

Using our techniques, we can provide the following more refined margin bound:

Theorem 16 *Under assumptions 1, 2, 3, let ℓ be a loss satisfying $\ell(z) = z^{-b}$ for all $z \geq 1$. Let $\mathbf{w}$ be some point reached by gradient descent, such that $\hat{L}(\mathbf{w}) = \epsilon$ for some $\epsilon \in (0, \ell(0))$. Then for any $p \in [\epsilon, 1]$, it holds that for at least $(1 - p)m$ of the indices $i \in [m]$ that*

$$\mathbf{x}_i^\top \bar{\mathbf{w}} \; > \; \frac{\gamma}{2} \cdot p^{1/b} \; .$$

Proof By definition, $\ell^{-1}(z) = z^{-1/b}$ for any $z \in (0, 1]$. Plugging this into Thm. 9, and noting that $\ell(0) \geq \ell(1) = 1$, it follows that for any $p \in [\epsilon, 1]$, for at least $(1 - p)m$ of the indices $i \in [m]$,

$$\mathbf{x}_i^\top \bar{\mathbf{w}} \; > \; \frac{\gamma}{2} \cdot \frac{(\epsilon/p)^{-1/b}}{\epsilon^{-1/b}} \; = \; \frac{\gamma}{2} \cdot p^{1/b} \; .$$

∎

Assuming gradient descent converges to a globally minimal value, we have $\epsilon \to 0$. As a result, after sufficiently many iterations, we can pick p slightly below $1/m$ (say $1/2m$), and get a margin bound of the form $\Omega(\gamma \cdot m^{-1/k})$ on all of the training examples (since getting a margin violation on less than $1/m$ of m examples is equivalent to no margin violations). This essentially recovers the result of Dudik et al. (2020) (up to the small difference of having b in the exponent instead of $b + 1$). However, our theorem covers a wider regime, since we can pick other values of p. For example, by picking p to be any constant, the theorem implies that we attain $\Omega(\gamma)$ margin on a constant portion of the training data (which can be arbitrarily close to 1).

Finally, we note that by upper bounding ϵ as a function of the number of iterations, the theorem above can be used to quantify how many iterations are required to achieve a certain margin on a certain percentage of the data points, as well as derive margin-based generalization bounds, in a manner completely analogous to the results in Sec. 4. Since polynomially-tailed losses are not commonly used in practice, we do not further pursue this here.

7. Numerical Experiment

We conclude our paper by providing a numerical experiment supporting our theoretical results on a real-world dataset.
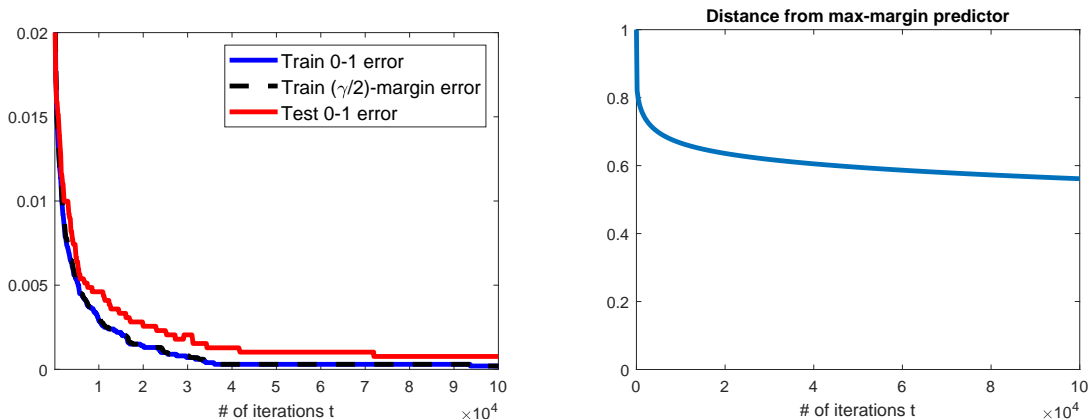
Figure 2: Running gradient descent on a linearly-separable subset of MNIST. Left figure: Train and test error, as a function of the number of iterations t. Right figure: Distance from the max-margin predictor $\mathbf{w}^*$, as a function of t. Best viewed in color.

Specifically, we used a linearly-separable subset of the well-known MNIST digit recognition dataset, focusing on the binary classification task of distinguishing between the letters 4 and 7 using a linear predictor. This dataset was constructed as follows: We took all 14117 examples of $\{4, 7\}$ digits in the MNIST dataset, centered and rescaled to have norm at most 1. Then, we trained an unbiased linear classifier using gradient descent (with step size 1 and using the logistic loss) until all but 212 examples were misclassified. After removing these examples, we were left with a dataset of 13905 examples, which by definition is linearly separable. Using a quadratic programming solver, we computed the exact max-margin predictor $\mathbf{w}^*$, with a margin of $\gamma := 0.0187...$

Given this linearly-separable dataset, we randomly split it into a training set of size 10^4, and a test set. We again ran gradient descent (with the logistic loss and step size of 1) on the training set. For each iterate, we computed the average $0-1$ loss and the average $\gamma/2$-margin error over the training set (that is, the percentage of training examples $(\mathbf{x}, y)$ where $y\mathbf{x}^\top \mathbf{w}_t \leq 0$ and $y\mathbf{x}^\top \mathbf{w}_t \leq \gamma/2$ respectively). We also computed the average $0-1$ loss over the test set, as well as the distance to the max-margin predictor $\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|$.

The results are presented in Fig. 7, from which several conclusions may be drawn:

- First, the train error (both $0-1$ and margin variants) go to zero with the number of iterations t, as should be expected (cf. Corollary 10).

- Second, we see that the test $0-1$ error decays with the training error, and actually remains quite close to it, appearing to converge to a small positive value. This is in complete agreement with the qualitative behavior predicted by our Thm. 11, where the test error remains comparable to the test error up to an unavoidable positive statistical term. In other words, at no point does gradient descent significantly overfit, regardless of the number of iterations.

15

- Third, the difference between the $0 - 1$ training error and the $\gamma/2$-margin error is negligible. This provides an indication that once an example is classified correctly, achieving a margin of $\Omega(\gamma)$ on it does not require significantly more iterations, on average (at least for this dataset whose margin is rather small, as is typical of real-world datasets). This lends support to our argument in Sec. 5 that the α parameter in Corollary 10 can be taken to be close to 1.

- Fourth, our results substantiate the claim in the introduction, that the overfitting behavior of gradient descent is dramatically better than what can be predicted solely based on convergence to the max-margin predictor $\mathbf{w}^*$: Indeed, the normalized iterates $\bar{\mathbf{w}}_t$ converge to $\mathbf{w}^*$ very slowly, and even after 10^5 iterations on this simple dataset, we have $\|\bar{\mathbf{w}}_t - \mathbf{w}^*\| \geq 0.56$. In contrast, to guarantee small training error based on distance from $\mathbf{w}^*$ alone, we must have $\|\bar{\mathbf{w}}_t - \mathbf{w}^*\| < \gamma = 0.0187...$

Finally, we note that since our theoretical bounds are worst-case over all datasets, they do not necessarily match the error curves quantitatively on this dataset. However, there exist datasets where they are quite tight up to constants, as evidenced by the constructions formally analyzed in Thm. 14 (see also Figure 1 for a graphical illustration).

Acknowledgments

Appendix A. Additional Proofs

A.1 Proof of Lemma 7

We first argue that $\hat{L}(\mathbf{w}_t)$ is monotonically decreasing in t. To see this, note that by the assumption that the derivative of ℓ (and hence the gradient of $\hat{L}$) is μ-Lipschitz, we have

$$
\begin{aligned}
\hat{L}(\mathbf{w}_{t+1}) &= \hat{L}(\mathbf{w}_t - \eta_t \nabla \hat{L}(\mathbf{w}_t)) \leq \hat{L}(\mathbf{w}_t) - \eta_t \|\nabla \hat{L}(\mathbf{w}_t)\|^2 + \frac{\mu}{2}\eta_t^2 \|\nabla \hat{L}(\mathbf{w}_t)\|^2 \\
&= \hat{L}(\mathbf{w}_t) - \eta_t \|\nabla \hat{L}(\mathbf{w}_t)\|^2 \left(1 - \frac{\mu}{2}\eta_t\right) ,
\end{aligned}
$$

which implies that $\hat{L}(\mathbf{w}_{t+1}) \leq \hat{L}(\mathbf{w}_t)$ whenever $\eta_t \leq 2/\mu$, which is indeed assumed. This also implies that $\hat{L}(\mathbf{w}^*) \leq \hat{L}(\mathbf{w}_T) \leq \hat{L}(\mathbf{w}_t)$ for all t.

Next, we argue that $\|\mathbf{w}_t - \mathbf{w}^*\|$ is monotonically decreasing in t. To see this, note that by smoothness of $\hat{L}$ and the monotonicity property above, for any $t < T$,

$$
\begin{aligned}
\hat{L}(\mathbf{w}^*) - \hat{L}(\mathbf{w}_t) &\leq \hat{L}(\mathbf{w}_T) - \hat{L}(\mathbf{w}_t) \leq \hat{L}(\mathbf{w}_{t+1}) - \hat{L}(\mathbf{w}_t) = \hat{L}(\mathbf{w}_t - \eta_t \nabla \hat{L}(\mathbf{w}_t)) - \hat{L}(\mathbf{w}_t) \\
&\leq -\eta_t \|\nabla \hat{L}(\mathbf{w}_t)\|^2 + \frac{\mu}{2}\eta_t^2 \|\nabla \hat{L}(\mathbf{w}_t)\| = -\eta_t \left(1 - \frac{\eta_t \mu}{2}\right) \|\nabla \hat{L}(\mathbf{w}_t)\|^2 \\
&\leq -\frac{\eta_t}{2}\|\nabla \hat{L}(\mathbf{w}_t)\|^2 ,
\end{aligned}
$$

where in the last step we used the fact that $\eta_t \leq 1/\mu$, hence $1 - \frac{\eta_t \mu}{2} \leq \frac{1}{2}$. Overall, we get that

$$
\|\nabla \hat{L}(\mathbf{w}_t)\|^2 \leq \frac{2}{\eta_t}(\hat{L}(\mathbf{w}_t) - \hat{L}(\mathbf{w}^*)) .
$$

Employing this inequality, together with $\hat{L}(\mathbf{w}_t) - \hat{L}(\mathbf{w}^*) \leq \nabla \hat{L}(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}^*)$ (which follows from convexity of $\hat{L}$), we have the following:

$$
\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 - \|\mathbf{w}_t - \mathbf{w}^*\| &= \|\mathbf{w}_t - \eta_t \nabla \hat{L}(\mathbf{w}_t) - \mathbf{w}^*\|^2 - \|\mathbf{w}_t - \mathbf{w}^*\| \\
&= -2\eta_t \nabla \hat{L}(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}^*) + \eta_t^2 \|\nabla \hat{L}(\mathbf{w}_t)\|^2 \\
&\leq -2\eta_t(\hat{L}(\mathbf{w}_t) - \hat{L}(\mathbf{w}^*)) + \eta_t^2 \cdot \frac{2}{\eta_t}(\hat{L}(\mathbf{w}_t) - \hat{L}(\mathbf{w}^*)) = 0 .
\end{aligned}
$$

To complete the proof, note that since $\|\mathbf{w}_t - \mathbf{w}^*\|$ is monotonically decreasing, then for any t, it is at most $\|\mathbf{w}_1 - \mathbf{w}^*\| = \|\mathbf{w}^*\|$. Therefore, $\|\mathbf{w}_t\| \leq \|\mathbf{w}_t - \mathbf{w}^*\| + \|\mathbf{w}^*\| \leq 2\|\mathbf{w}^*\|$.

A.2 Proof of Corollary 10

By Lemma 8, we know that at iteration T, $\hat{L}(\mathbf{w}(T)) = \epsilon$ for some

$$
\epsilon \leq \frac{1}{T} + \frac{\log^2(T)}{2\gamma^2 T} \leq \frac{1}{\gamma^2 T}\left(1 + \frac{1}{2}\log^2(T)\right) \leq \frac{\log^2(T)}{\gamma^2 T} ,
$$

where we used the facts that γ must be at most 1 and $1 \leq \frac{1}{2}\log^2(T)$ (since $T > 4$). By the theorem assumptions, it follows that $\epsilon < \ell(0)$, so we must have $\mathbf{w}(T) \neq \mathbf{0}$. Applying Thm. 9 (noting that ℓ has $\frac{1}{4}$-Lipschitz gradients and that $\ell^{-1}(z) := \log(1/(\exp(z) - 1))$,

17

which is between $\log(1/z)$ and $\log(1/2z)$ for all $z \in [0,1]$), we get that for any $p \in [\frac{\epsilon}{\ell(0)}, 1]$, for at least $(1-p)m$ of the indices i,

$$\mathbf{x}_i^\top \bar{\mathbf{w}}_T > \frac{\gamma}{2} \cdot \frac{\log(p/2\epsilon)}{\log(1/\epsilon)} = \frac{\gamma}{2} \cdot \left(1 - \frac{\log(2/p)}{\log(1/\epsilon)}\right) .$$

In particular, picking $p = 2\epsilon^\alpha$ for some $\alpha \in [0,1]$ (which satisfies $p \geq \frac{\epsilon}{\ell(0)} = \frac{\epsilon}{\log(2)}$), and noting that $p \leq 2\left(\frac{\log^2(T)}{\gamma^2 T}\right)^\alpha$, the result follows.

A.3 Proof of Lemma 15

We will utilize the following easily-verified facts about the logistic loss $\ell(z) = \log(1 + \exp(-z))$:

$$\forall z \in \mathbb{R} , \quad -1 \leq \ell'(z) < 0$$
$$\forall z \leq 1 , \quad \ell'(z) < -\frac{1}{4} \tag{6}$$

Also, by definition of gradient descent, we have the following:

$$r_1 = s_1 = 0$$
$$r_{t+1} = r_t - (1-\epsilon)\eta \cdot \ell'(r_t) + \frac{\epsilon\eta}{2} \cdot \ell'\left(-\frac{1}{2}r_t + 3\gamma s_t\right)$$
$$s_{t+1} = s_t - 3\gamma\epsilon\eta \cdot \ell'\left(-\frac{1}{2}r_t + 3\gamma s_t\right) . \tag{7}$$

The proof ofrelies on the following key lemma:

Lemma 17 *There exists an iteration index $T \leq \frac{32}{5\eta} + 1$ for which*

1. *$r_t \geq \frac{15}{16}$ for all $t \geq T$*

2. *$s_T \leq \frac{3}{10}$*

3. *$-\frac{1}{2}r_t + 3\gamma s_t \leq 0$ for all $t \leq T$.*

Proof We start with the first item. If $r_t \leq 1$, then by Eq. (6), Eq. (7) and the assumption on ϵ,

$$r_{t+1} \geq r_t + (1-\epsilon)\eta \cdot \frac{1}{4} - \frac{\epsilon\eta}{2} = r_t + \frac{\eta}{4}(1 - 3\epsilon) \geq r_t + \frac{5\eta}{32} .$$

Since we start at $r_1 = 0$, we get that as long as $r_t \leq 1$, r_t increases in increments of at least $5\eta/32$. Thus, there is some index $T \leq \frac{32}{5\eta} + 1$ at which $r_T \geq 1$ for the first time, and r_t is monotonically increasing for all $t \leq T$. Once $r_t \geq 1$, it cannot decrease to $\frac{15}{16}$ or below in the following iteration, since by the update equation,

$$r_{t+1} \geq r_t + 0 - \frac{\epsilon\eta}{2} \geq r_t - \frac{1}{16} \geq \frac{15}{16} ,$$

and if $r_{t+1} \leq 1$, it must monotonically increase again as shown earlier. This establishes the first item in the lemma.

18

Turning to the second item, we note that $s_1 = 0$ and for any t, $s_{t+1} \leq s_t + 3\gamma\epsilon\eta \leq s_t + 3 \cdot \frac{1}{8} \cdot \frac{1}{8} \cdot \eta = s_t + \frac{3\eta}{64}$. Since $T \leq \frac{32}{5\eta} + 1$, it follows that

$$s_T \ \leq \ \frac{3\eta}{64}(T-1) \ \leq \ \frac{3\eta}{64} \cdot \frac{32}{5\eta} \ = \ \frac{3}{10} \ .$$

Turning to the third item in the lemma, define for simplicity $u_t := -\frac{1}{2}r_t + 3\gamma s_t$. Thus, we need to show $u_t \leq 0$ for all $t \leq T$. This trivially holds for $t = 1$. For any $t < T$, by the update equations for r_t, s_t,

$$
\begin{aligned}
u_{t+1} \ &= \ u_t + \frac{(1-\epsilon)\eta}{2} \cdot \ell'(r_t) - \left(\frac{1}{4} + 9\gamma^2\right)\epsilon\eta \cdot \ell'(u_t) \\
&\overset{(*)}{\leq} \ u_t - \frac{(1-\epsilon)\eta}{8} + \left(\frac{1}{4} + \frac{9}{64}\right)\epsilon\eta \ = \ u_t - \frac{\eta}{8}\left(1 - \frac{33}{8}\epsilon\right) \ \overset{(**)}{<} \ u_t \ .
\end{aligned}
$$

where in $(**)$ we used the assumption $\epsilon \leq \frac{1}{8}$, and in $(*)$ we used the facts that $\gamma \leq \frac{1}{8}$, that $-1 \leq \ell'(z) \leq 0$ for any z, and that since $t < T$, $r_t < 1$ (see the definition of T above), hence $\ell'(r_t) < -\frac{1}{4}$. Overall, we get that u_t is monotonically decreasing for all $t \leq T$. But we have $u_1 = -\frac{1}{2}r_1 + 3\gamma s_1 = 0$, hence u_t must be non-positive for all $t \leq T$ as required. ∎

With this lemma at hand, we turn to prove Lemma 15. According to the lemma, $-\frac{1}{2}r_t + 3\gamma s_t > 0$ can only occur for $t > T$ (in which case, $r_t \geq \frac{15}{16}$). This requires that

$$s_t \ > \ \frac{1}{6\gamma}r_t \ \geq \ \frac{15}{96\gamma} \ .$$

However, by the lemma, we have $s_T \leq \frac{3}{10}$, and by the update rule for s_t, $s_{t+1} \leq s_t + 3\gamma\epsilon\eta \leq s_t + 3\gamma\epsilon$. Thus, the number of additional iterations (after iteration T) required to make $s_t > \frac{15}{96\gamma}$ is at least $\frac{15/96\gamma - 3/10}{3\gamma\epsilon} \ = \ \frac{5}{96\gamma^2\epsilon}\left(1 - \frac{96}{50}\gamma\right) \geq \frac{5}{96\gamma^2\epsilon}\left(1 - \frac{96}{50\cdot8}\right) = \frac{19}{480\gamma^2\epsilon}$ as required.

References

Miroslav Dudik, Ziwei Ji, Robert Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, 2020.

Suriya Gunasekar. Characterizing implicit bias in terms of optimization geometry. In *In International Conference on Machine Learning*, 2018.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

Steve Hanneke and Aryeh Kontorovich. Optimality of svm: Novel proofs and tighter bounds. *Theoretical Computer Science*, 796:99–113, 2019.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2018a.

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

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019a.

Ziwei Ji and Matus Telgarsky. A refined primal-dual analysis of the implicit bias. *arXiv preprint arXiv:1906.04540*, 2019b.

David McAllester. Simplified pac-bayesian margin bounds. In *Learning theory and Kernel machines*, pages 203–215. Springer, 2003.

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, 2019a.

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, 2019b.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

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.