THE IMPLICIT BIAS OF GRADIENT DESCENT: FROM LINEAR CLASSIFIERS TO DEEP NETWORKS

BY

ZIWEI JI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Assistant Professor Matus Telgarsky, Chair
Professor Chandra Chekuri
Associate Professor Maxim Raginsky
Associate Professor Daniel Hsu, Columbia University

Abstract

Gradient descent and other first-order methods have been extensively used in deep learning. They can find solutions with not only low training error, but also low test error. This motivates the study of the *implicit bias* of training algorithms such as gradient descent, meaning we want to understand what special properties are satisfied by gradient descent solutions that lead to good generalization. In this thesis, we focus on gradient descent and its derivatives, show test error bounds and characterize the implicit biases.

In detail, for linear classifiers, we consider both the separable setting and nonseparable setting. In the separable case, we first give an $1/t$ test error bound for stochastic gradient descent. Then for gradient descent, we present a primal-dual framework to analyze its implicit bias, which leads to fast margin maximization algorithms. While the previous results mostly require exponentially-tailed losses, we also show that for general decreasing losses, the implicit bias can still be characterized in terms of the regularization path. In the nonseparable case, we design a nearly-optimal algorithm by combining logistic regression and perceptron. We also characterize the implicit bias of gradient descent via a unique decomposition of the training set.

For neural networks, we first provide test error bounds for shallow ReLU networks, using the recent idea of neural tangent kernel, but only requiring a mild overparameterization. Then we show implicit bias results for deep linear networks and deep homogeneous networks, in the form of *alignment* properties.

Acknowledgments

First, I want to thank my advisor Matus Telgarsky. He gave me lots of guidance and support, from high-level research tastes to detailed techniques, which are extremely helpful to me.

I was also fortunate to be advised by Ruta Mehta during my first year. She led me through my first PhD project, which is very important for my PhD study.

I did two internships during my PhD, and I want to thank my mentors: Miro Dudík and Robert Schapire at Microsoft Research NYC, and Pranjal Awasthi and Satyen Kale at Google Research NYC. I learned many different techniques and research topics during the internships, which are very valuable to me.

I am also grateful to the long list of people who have given advice or collaborated with me: Kwangjun Ahn, Bolton Bailey, Niladri Chatterji, Chandra Chekuri, David Forsyth, Daniel Hsu, Yuzheng Hu, Prateek Jain, Nan Jiang, Stefani Karp, Sanmi Koyejo, Justin Li, Zhiyuan Li, Kaifeng Lyu, Praneeth Netrapalli, Maxim Raginsky, Gil Shamir, Nathan Srebro, Belinda Tzen, Lan Wang, and Ruicheng Xian.

Finally, I am grateful to my parents; they always believe in me and are proud of me. I also thank Zixin Huang for her support during my PhD study.

Table of Contents

Chapter 1: Introduction

In recent years, deep learning has achieved great empirical success in many areas, such as computer vision (e.g., ResNet [1]), natural language processing (e.g., Transformer [2]), and deep reinforcement learning (e.g., AlphaGo [3]). However, there are still fundamental questions not fully answered: *why are optimization and generalization feasible in deep learning?* To give some background, we first introduce the notions of optimization and generalization on a simple but typical setting in machine learning: suppose we have a training set $\{(x_i, y_i)\}_{i=1}^n$ where each x_i represents an image of either a table or chair, and $y_i \in \{-1, +1\}$ represents the label where -1 denotes a table and $+1$ denotes a chair. Now the high-level goal of machine learning is to learn useful patterns from the training set, which further consists of *optimization* and *generalization*. The problem of optimization means that we want to train a model or function f, such that for most training examples, it holds that $f(x_i) = y_i$. On the other hand, we also need to consider generalization, meaning we also want our model to make good predictions on some unseen data. For example, if we train a model on a training set of tables and chairs, we also want our model to be able to distinguish tables from chairs on some future inputs.

Optimization is well-understood for convex objective functions; on the other hand, there exists a simple nonconvex function that requires exponential time to optimize [4, Theorem 1.1.2]. In deep learning, the training objective is usually highly nonconvex, but gradient-based algorithms such as stochastic gradient descent (SGD) [5] and Adam [6] can usually still obtain a high training accuracy despite nonconvexity. In recent years, a lot of progress has been made to explain the feasibility of training in deep learning, but these prior results also have different kinds of limitation, and a deeper analysis is still required.

Generalization for neural networks has also been studied for a long time. However, recently people have noticed that existing results are not enough to explain practical success of deep networks. For example, a classical tool to analyze generalization is the VC-dimension, which has also been applied to neural networks [7]. Specifically, if the network can only represent a limited number of sign patterns, then the VC-dimension analysis can rigorously ensure good generalization. However, it has been found in practice that neural networks can even fit random signs [8], suggesting that the VC theory may not be enough to explain the strong generalization performance of neural networks in practice, and thus a more fine-grained analysis is needed.

In this thesis, we try to answer the above questions by analyzing optimization and generalization *simultaneously*. The motivation is that, certain training algorithms such as SGD can

usually find a solution that can not only fit the training data, but also have good generalization [8]. Motivated by this, we try to study the *implicit bias* of training algorithms including SGD, i.e., to characterize some special properties of the SGD solution, which may further be used to prove good generalization bounds and make the model more interpretable. On the other hand, as we will see below, the study of implicit bias also leads to finer optimization analyses, which can be useful in understanding the feasibility of training in deep learning.

In the remaining parts of the introduction, we first introduce the problem setup and some common notation, and then give a summary of our results that will be discussed in this thesis. Detailed theorem statements and proofs, along with comparisons with many related works, will be given in following sections.

1.1 NOTATION AND PROBLEM SETUP

In this section, we first introduce some general notation that will be used throughout this thesis, then give a formal description of the problem setup.

Let $\mathbb{1}_A$ denote the indicator of an event A, i.e., $\mathbb{1}_A = 1$ when A happens, and $\mathbb{1}_A = 0$ when A does not happen. Given a convex set C, let ι_C denote the indicator function, i.e., $\iota_C(x) = 0$ if $x \in C$, and $\iota_C(x) = \infty$ if $x \notin C$.

A real-valued differentiable function f is called β-smooth with respect to norm $\|\cdot\|$ if its gradient is β-Lipschitz continuous with respect to norm $\|\cdot\|$; formally, for any x, x' in its domain, we have $\big\|\nabla f(x) - \nabla f(x')\big\|_* \le \beta\|x - x'\|$, where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$. In particular, this implies [9, Lemma 3.4]

$$\Big|f(x') - f(x) - \big\langle \nabla f(x), x' - x\big\rangle\Big| \le \frac{\beta}{2}\|x - x'\|^2.$$

Next we give the problem setup. In this thesis, we consider a binary classification problem, as described below. Without loss of generality, we suppose the inputs or feature vectors are from $\mathbb{R}^d$, while the label is either -1 or $+1$, denoting two classes. For example, each input may be an image of a table or chair, and if the input is a table, its label is -1, otherwise its label is $+1$. In the following, we often deal with a training set $\{(x_i, y_i)\}_{i=1}^n$ which consists of n training examples, and we assume they are identically and independently sampled from an underlying data distribution P over $\mathbb{R}^d \times \{-1, +1\}$. Informally, our goal is to learn a model or function based on the training set, which can map the input feature vector to the correct label as much as possible, on both the training set and the distribution P. Below we give a formal description of this problem.

The model will be denoted by a real-valued function $f(x; w)$ in general, where x denotes the

input feature vector and w denotes the weight or parameter vector. Given a feature vector x, to make a prediction, we simply use the sign of the output of the model sign $\big(f(x;w)\big)$. In this thesis, we will consider either a linear classifier $f_{\text{lin}}(x;w) := \langle w, x\rangle$, or a neural network as described below. An L-layer neural network consists of the following parts: (i) L weight matrices $W_L, \dots, W_1$, where $W_k \in \mathbb{R}^{d_k \times d_{k-1}}$; (ii) L bias vectors $b_L, \dots, b_1$, where $b_k \in \mathbb{R}^{d_k}$; (iii) an activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Given an input $x \in \mathbb{R}^{d_0}$, let $x_0 := 0$, then for each $1 \le k \le L-1$, the neural network computes $x_k := \sigma(W_k x_{k-1} + b_k)$, where σ is applied to each coordinate. The final output is given by $W_L x_{L-1} + b_L$. The network defined above can be described compactly as below:

$$f_{\text{nn}}(x;w) := W_L \sigma\Big(W_{L-1}\sigma\big(\cdots\sigma(W_1 x + b_1)\cdots\big) + b_{L-1}\Big) + b_L,$$

where w includes all parameters $W_1, \dots, W_L, b_1, \dots, b_L$.

We will focus on the *risk minimization* framework. Given a decreasing loss function $\ell : \mathbb{R} \to \mathbb{R}$, the population risk and empirical risk of w are given by

$$\mathcal{R}(w) := \mathbb{E}_{(x,y)\sim P}\Big[\ell\big(yf(x;w)\big)\Big], \quad \text{and} \quad \widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i f(x_i;w)\big). \tag{1.1}$$

The following quantities will also be useful in our analysis:

$$\mathcal{Q}(w) := \mathbb{E}_{(x,y)\sim P}\Big[-\ell'\big(yf(x;w)\big)\Big], \quad \text{and} \quad \widehat{\mathcal{Q}}(w) := \frac{1}{n}\sum_{i=1}^{n} -\ell'\big(y_i f(x_i;w)\big), \tag{1.2}$$

where ℓ' denotes the derivative of ℓ. The problem of optimization is usually just to minimize $\widehat{\mathcal{R}}$; note that since ℓ is decreasing, by minimizing $\widehat{\mathcal{R}}$, we effectively try to force y_i and $f(x_i)$ to have the same sign. On the other hand, to show that our model also generalizes well, we will prove upper bounds on $\mathcal{R}$.

Here are some examples of decreasing loss functions:

- The logistic loss $\ell_{\log}(z) := \ln(1 + e^{-z})$: widely-used in practice.
- The hinge loss $\ell_{\text{h}}(z) := \max\{-z + 1, 0\}$: also widely-used in practice.
- The exponential loss $\ell_{\exp}(z) := e^{-z}$: this one is rarely used in practice, but it will allow for a clean implicit-bias analysis.
- The ReLU loss $\ell_{\text{r}}(z) := \max\{-z, 0\}$: this one is sometimes also called the hinge loss; we call it the ReLU loss to distinguish it from the previous one. Note that with the ReLU

loss, 0 is a global minimizer of the risk function, and it may seem that no optimization is needed. However, by letting $\ell_{\mathrm{r}}'(0) = -1$ (cf. Section 2.1) or by choosing a special domain (cf. Section 3.1), we can also obtain efficient algorithms with the ReLU loss.

Moreover, let $\ell_{0-1}(z) := \mathbb{1}_{z \le 0}$ denote the zero-one loss, we will often try to derive bounds on the population zero-one risk, or test misclassification error

$$\mathcal{R}_{0-1}(w) := \mathbb{E}_{(x,y)\sim P}\left[\ell_{0-1}\left(yf(x;w)\right)\right] = \Pr_{(x,y)\sim P}\left(y \neq \operatorname{sign}\left(f(x;w)\right)\right).$$

The most basic algorithms we will study are gradient descent (GD) and stochastic gradient descent (SGD). For both algorithms, we start from some properly-chosen initialization w_0. For GD, we will work with a training set and the corresponding empirical risk, and let

$$w_{t+1} := \Pi_D\left[w_t - \eta_t \nabla \widehat{\mathcal{R}}(w_t)\right],$$

where η_t is the learning rate or step size at step t, and D is the domain of optimization. In most cases in this thesis, we simply let $D = \mathbb{R}^d$ (i.e., no projection), but in certain cases a carefully-chosen D can allow for an efficient algorithm (cf. Section 3.1). For GD, we can either prove a bound on the empirical risk $\widehat{\mathcal{R}}(w_t)$, or study if the GD iterates w_t converge to a global minimizer, or even prove a test misclassification error bound $\mathcal{R}_{0-1}(w_t)$ using a generalization analysis. For SGD, we sample a new data example (x_t, y_t) at step t, and let

$$w_{t+1} := \Pi_D\left[w_t - \eta_t \ell'\left(y_t f(x_t; w_t)\right) \nabla_w f(x_t; w_t)\right].$$

We usually directly prove a test error bound $\mathcal{R}_{0-1}(w_t)$ for SGD. Based on our understanding of GD and SGD, we will also design more efficient algorithms in different settings (cf. Sections 2.3 and 3.1).

Sometimes we also consider gradient flow, defined as a solution to the differential equation

$$\frac{\mathrm{d}w_t}{\mathrm{d}t} = -\nabla \widehat{\mathcal{R}}(w_t).$$

In other words, gradient flow can be viewed as gradient descent with infinitesimal learning rates. A gradient flow may not exist if $\widehat{\mathcal{R}}$ is not smooth enough, but in this thesis we will not spend too much effort on such technical details; instead, we use gradient flow mainly to simplify the analysis and highlight the key proof ideas, particularly when dealing with deep networks. Moreover, many of our results can also be extended to the gradient descent case with sufficiently small learning rates.

1.2 SUMMARY OF RESULTS

In this section, we give a brief summary of the results that will be covered in this thesis. We will consider both linear classifiers and neural networks, and try to show test error bounds and characterize implicit biases.

1.2.1 Linear classifiers

For simplicity, we start our analysis from linear classifiers: the input x is mapped to $\langle w, x\rangle$, where w denotes the weight vector. Linear classifiers have been studied for a long time, but we still provide some novel results and analyses as detailed below; moreover, these ideas can also be useful when analyzing neural networks.

Roughly speaking, we will consider two cases: the separable case and nonseparable case. We discuss them separately below.

Separable case. The separability assumption can be made on either the training set or the data distribution P. Given a training set $\{(x_i, y_i)\}_{i=1}^n$, separability means that there exists a linear classifier u such that $y_i\langle u, x_i\rangle > 0$ for all $1 \le i \le n$. In other words, there exists a ground-truth solution that can give the correct label on all training examples. Similarly, given the underlying data distribution P, we can assume a stronger separability condition: there exists u such that for almost all $(x, y) \sim P$, it holds that $y\langle u, x\rangle > 0$. In this thesis, we will give test error bounds under the distributional separability assumption, and further give finer characterizations of the implicit bias of GD with a separable training set.

A central notion in the analysis of separable case is *margin*: with a separable training set, given a ground-truth solution u, its margin γ is given by $\gamma := \min_{1\le i\le n} y_i\langle u, x_i\rangle > 0$. Similarly, in the distributional setting, the exact separability assumption we will make also includes a concrete margin: there exists u and $\gamma > 0$ such that $y\langle u, x\rangle \ge \gamma$ almost surely. Intuitively, after a proper normalization (e.g., after ensuring $\|u\|_2 = 1$), the larger the margin is, the easier the problem is; indeed, margin has been used in the design and analysis of many machine learning algorithms, such as Perceptron [10], SVM [11], Boosting [12], and neural networks [13].

Chapter 2 focuses on linear classifiers under the separable case. First, if the data distribution P can be separated with margin γ, Theorem 2.1 proves an $\widetilde{O}\left(\frac{1}{\gamma^2 t}\right)$ test misclassification error bound for SGD with the logistic loss, based on results from [14]. The analysis can handle constant learning rate, thus giving the $\widetilde{O}(1/t)$ test error bound. The only prior $\widetilde{O}(1/t)$ rate requires strong convexity [15], which unfortunately does not hold in our setting: the rea-

son is that we consider an unbounded domain $\mathbb{R}^d$ on which the risk function is not strongly convex, since we use the logistic loss which is not strongly convex on $\mathbb{R}$.

On the other hand, if we focus on a linearly separable training set, we can provide finer characterizations of the implicit bias of GD: more exactly, we want to know what properties are satisfied by GD solutions that lead to good generalization, and if possible, we also want to characterize the limit of the GD iterates. Previously, [16] proved that GD with the logistic loss converges to the ℓ_2 maximum-margin solution. In a parallel work [17], we also show margin maximization using a different proof technique based on [18]. However, these prior results do not give the fastest possible rate. In Section 2.2, we develop a primal-dual framework to analyze the implicit bias based on [19]: we show that GD on the original problem induces a mirror descent update on a dual problem, whose optimum is exactly given by the maximum margin, thus giving a natural interpretation of the implicit bias phenomenon. This framework also allows us to give faster analyses and algorithms for the following margin maximization problem:

$$\max_{\|w\|_2 \le 1} \min_{1 \le i \le n} y_i \langle w, x_i \rangle. \tag{1.3}$$

First of all, we show a fast $O(1/t)$ margin maximization rate for GD (cf. Theorem 2.3). Furthermore, motivated by the primal-dual framework, we design a momentum-style margin maximization algorithm that has an even faster $\widetilde{O}(1/t^2)$ rate (cf. Theorem 2.5), which is based on [20]. Prior first-order methods mostly have rate $O(1/\sqrt{t})$; for example, we can apply gradient ascent to eq. (1.3) (we use ascent since we need to maximize a concave function); this is the batch perceptron algorithm [21], and its rate is $O(1/\sqrt{t})$. On the other hand, we may also apply the ellipsoid method to eq. (1.3); to ensure an additive error of ϵ, the ellipsoid method needs $O(d^2 \ln(1/\epsilon))$ iterations, while our momentum-style method needs $O(1/\sqrt{\epsilon})$ iterations. Therefore, for $\epsilon \ge 1/\mathrm{poly}(d)$, our method is faster. More related work is discussed in Section 2.2.

The above implicit bias results all require a loss function with an exponential tail (e.g., the exponential loss or logistic loss); in Section 2.4, we further show that even for general decreasing losses, the implicit bias of GD can be characterized in terms of the regularization path (cf. Theorems 2.7 and 2.8). This section is based on [22].

Nonseparable case. Next we consider a general nonseparable setting.

In the distributional setting, we assume the optimal test misclassification error that can be obtained by linear classifiers is given by $\mathrm{OPT} > 0$, and we let $\bar{u}$ denote the optimal linear classifier. Our goal is to find a solution that can compete with $\bar{u}$. Previous state-of-

the-art algorithms can reach a test error bound of $C \cdot \mathrm{OPT}$ for some universal constant C [23, 24], but their algorithms are complicated. On the other hand, since logistic regression is one of the most fundamental algorithms in machine learning, it is natural to try to apply it to this problem; however, previously only an $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound was known [25]. In Section 3.1, we first provide a lower bound of $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ for logistic regression (cf. Theorem 3.1), thus matching the upper bound in [25]. Moreover, we also show that we can overcome this lower bound, simply by first running logistic regression and then applying the Perceptron algorithm with a restricted domain and a warm start; this algorithm is very simple, and can reach an $\widetilde{O}(\mathrm{OPT})$ upper bound (cf. Theorem 3.2). This section is based on [26].

Next in Section 3.2, we consider a general training set (i.e., we do not make any assumption on the training set) and characterize the implicit bias of GD. First, we show that the training set can be uniquely decomposed into two parts: a maximal linearly separable part, and the remaining part which induces a strongly-convex optimization problem (cf. Theorem 3.3). We also show that, with no prior knowledge of this decomposition, GD can find the correct implicit bias over the parts: on the strongly convex part, GD finds the unique optimizer, while on the maximal linearly separable part, GD converges in direction to the maximum-margin classifier that is orthogonal to the strongly convex part (cf. Theorem 3.5). This section is based on [17].

1.2.2 Neural networks

Next we turn to neural networks. In this thesis, we will focus on the case where the data can be separated by a neural network, since in practice neural networks can often achieve 100% training accuracies; however it is also interesting to consider the nonseparable case, and we have also started considering this setting in [27].

Two-layer ReLU networks. We start our analysis from (wide) two-layer ReLU networks, using the idea of the neural tangent kernel (NTK) [28]. Formally, let $W \in \mathbb{R}^{m \times d}$ denote the first weight matrix, and $a \in \mathbb{R}^m$ denote the second weight matrix, the network outputs $f(x; W) := a^\top \max\{Wx, 0\}$ where the max operator is applied to each coordinate. For simplicity, we fix a and only train W, but our ideas can also be applied to the deep case [29]. We also let $\nabla f(x; W) \in \mathbb{R}^{m \times d}$ denote the gradient of the network with respect to the weight matrix W.

The idea of an NTK analysis is that, if the network is wide enough, then (i) we can already fit the training data using the random features $\nabla f(x_i; W_0)$ given by a wide network at random

initialization W_0, and (ii) during the GD process, these features stay roughly unchanged, i.e., $\nabla f(x_i; W_t) \approx \nabla f(x_i; W_0)$. In other words, the analysis of GD on a nonlinear network can be approximated by the analysis of GD on a linear model (on a high-dimensional space $\mathbb{R}^{m \times d}$). The latter GD process is easier to analyze due to convexity, and the approximation error can be shown to be small if the network is wide enough. An empirical risk bound can be shown using the above framework, and it is also possible to derive a test error bound via a careful generalization analysis.

However, prior NTK results always require the width to be very large, such as a polynomial of the number of training examples or inverse target error. In Chapter 4, we show that actually a polylogarithmic width is enough for good training and test error bounds, using the corresponding NTK margin (cf. Theorems 4.1 to 4.3). Our analysis is centered upon the notion of NTK margin (cf. Assumptions 4.1 and 4.2), which is basically the linear margin in the space of gradients $\nabla f(x; W_0)$. The NTK margin also allows us to show a tight sample complexity bound (cf. Section 4.4). This two-layer analysis is based on [30].

Deep homogeneous networks. Next we consider deep homogeneous networks, meaning that given input x and model parameter w, if we scale w by a positive constant c, then the output of the network is scaled by c^L for some $L > 0$ (i.e., $f(x; cw) = c^L f(x; w)$). Examples include deep networks with linear and convolutional layers, max and average pooling layers, and homogeneous activations, such as the identity activation $x \mapsto x$, ReLU activation $x \mapsto \max\{0, x\}$, and more generally powers of ReLU $x \mapsto \max\{0, x\}^k$. On the other hand, homogeneity does not allow skip connections and bias vectors. Here is a typical homogeneous network, where the activation σ is homogeneous:

$$x \mapsto W_L \sigma \left(W_{L-1} \sigma \left(\cdots \sigma(W_1 x) \cdots \right) \right).$$

The simplest homogeneous network is the deep linear network, which maps the input x to $W_L W_{L-1} \cdots W_2 W_1 x$, where $W_L, \ldots, W_1$ are the weight matrices. A deep linear network has poor expressive power, since it can only represent a linear function; on the other hand, it still induces a nonconvex optimization problem, and an analysis of it may shed some light on nonlinear networks. In Section 5.1, we show that despite overparameterization and nonconvexity, gradient flow can find a very simple solution: all weight matrices become nearly rank-1, adjacent weight matrices tend to have identical top singular vectors, and the whole network computes the maximum-margin predictor (cf. Theorems 5.1 and 5.2). This section is based on [31].

Then in Section 5.2, we further generalize the previous result to deep homogeneous net-

works. We show that the gradient flow iterate w_t and the corresponding (negative) gradient $-\nabla\widehat{\mathcal{R}}(w_t)$ converge to the same direction (cf. Theorem 5.3), meaning

$$\lim_{t\to\infty} \left\langle \frac{w_t}{\|w_t\|_F}, \frac{-\nabla\widehat{\mathcal{R}}(w_t)}{\left\|\nabla\widehat{\mathcal{R}}(w_t)\right\|_F} \right\rangle = 1.$$

This result is from [32]; it implies the previous result for deep linear networks, and can also be applied in many other settings.

Chapter 2: Linear classifiers with linearly-separable data

In this chapter, we focus on linear classifiers, and also assume the data is linearly separable with a positive margin. Formally, with the underlying data distribution P, we assume there exists a unit linear classifier u and a positive constant γ such that $y\langle u, x\rangle \geq \gamma$ almost surely. Alternatively, we may only focus on the training set, and assume $y_i\langle u, x_i\rangle \geq \gamma$ for all $1 \leq i \leq n$. Additionally, let

$$u^* := \arg\max_{\|u\|_2=1} \min_{1\leq i\leq n} y_i\langle u, x_i\rangle \quad \text{and} \quad \gamma^* := \max_{\|u\|_2=1} \min_{1\leq i\leq n} y_i\langle u, x_i\rangle$$

denote the maximum-margin classifier and maximum margin, respectively. The existence and uniqueness of u^* is guaranteed by Lemma 2.3, and thus its definition is valid. For simplicity, we will also assume $\|x\|_2 \leq 1$ almost surely (in the distributional case), or $\|x_i\|_2 \leq 1$ for all $1 \leq i \leq n$ (with a training set).

The Perceptron algorithm [33] is a classical algorithm to solve this problem, and Novikoff [10] showed an $O(1/\gamma^2)$ test error bound under the linear separability assumption. In Section 2.1, we will first review the proof of this result, and then show that it can be adapted to give a clean generalization analysis for SGD with the logistic loss (cf. Theorem 2.1). Similar ideas will also be useful in the nonseparable setting (cf. Chapter 3) and in the analysis of shallow ReLU networks (cf. Chapter 4).

In addition to a generalization bound, it is actually possible to give a much finer characterization of the solution found by GD. Previously, Soudry et al. [16] showed that if we run GD on a linearly separable training set with certain exponentially-tailed losses including the exponential loss and logistic loss, then $\|w_t\|_2 \to \infty$ while $w_t/\|w_t\|_2 \to u^*$. In other words, if we keep running GD with logistic regression, then it actually finds the same solution as the hard-margin support vector machine [11]. In a parallel work [17], we also proved this result, using a technique from [18]. However, these approaches are not able to give the fastest possible convergence rates. In Section 2.2, we will show that GD actually induces a mirror descent update on a certain dual problem whose optimum is exactly given by the maximum margin. This observation gives a clean interpretation of the implicit bias phenomenon, and allows us to give a fast $O(1/t)$ margin maximization rate (cf. Theorem 2.3).

In Section 2.3, we will further exploit the primal-dual framework and design a fast margin-maximization algorithm that achieves an $\widetilde{O}(1/t^2)$ rate (cf. Theorem 2.5). The key observation is that the dual objective is smooth, and thus can be optimized by Nesterov's accelerated method [4, 34, 35]. Moreover, the dual Nesterov method can be transformed into a primal

momentum method, for which we prove the $\widetilde{O}(1/t^2)$ rate.

Finally, above results all require certain exponentially-tailed losses. In Section 2.4, we will show that even for general decreasing losses (e.g., a loss with a polynomial tail), the implicit bias of GD can still be characterized in terms of the regularization path, which in general is different from the maximum-margin solution (cf. Theorems 2.7 and 2.8).

Finally, we discuss related work in Section 2.5.

2.1 SGD TEST ERROR BOUNDS

Recall that $\ell_{\mathrm{r}}(z) := \max\{-z, 0\}$. Starting from $w_0 := 0$, we consider SGD with ℓ_{r}:

$$w_{t+1} := w_t - \ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right) y_t x_t.$$

More explicitly, if $y_t\langle w_t, x_t\rangle > 0$, then $w_{t+1} = w_t$, otherwise $w_{t+1} = w_t + y_t x_t$ (recall that we let $\ell_{\mathrm{r}}'(0) = -1$). This is the *Perceptron* algorithm [33]. For simplicity, let $z_t := y_t x_t$, and thereby the update can be written as

$$w_{t+1} := w_t - \ell_{\mathrm{r}}'\left(\langle w_t, z_t\rangle\right) z_t.$$

Below is a standard analysis from [10]: note that

$$\begin{aligned}
\|w_{t+1}\|_2^2 &= \|w_t\|_2^2 - 2\left\langle \ell_{\mathrm{r}}'\left(\langle w_t, z_t\rangle\right) z_t, w_t\right\rangle + \ell_{\mathrm{r}}'\left(\langle w_t, z_t\rangle\right)^2 \|z_t\|_2^2 \\
&\le \|w_t\|_2^2 + 2\left(\ell_{\mathrm{r}}(0) - \ell_{\mathrm{r}}\left(\langle w_t, z_t\rangle\right)\right) + \ell_{\mathrm{r}}'\left(\langle w_t, z_t\rangle\right)^2 \\
&\le \|w_t\|_2^2 + \ell_{\mathrm{r}}'\left(\langle w_t, z_t\rangle\right)^2 = \|w_t\|_2^2 - \ell_{\mathrm{r}}'\left(\langle w_t, z_t\rangle\right),
\end{aligned}$$

where we use convexity and $\|x\|_2 \le 1$ in the first inequality, and that $\ell_{\mathrm{r}}(z) \ge \ell_{\mathrm{r}}(0) = 0$ in the second inequality, and that $(\ell_{\mathrm{r}}')^2 = -\ell_{\mathrm{r}}'$ in the end. It then follows that

$$\mathbb{E}\left[\|w_t\|_2^2\right] \le \mathbb{E}\left[\sum_{j<t} -\ell_{\mathrm{r}}'\left(\langle w_j, z_j\rangle\right)\right] = \mathbb{E}\left[\sum_{j<t} \mathcal{R}_{0-1}(w_j)\right],$$

since $-\ell_{\mathrm{r}}'(z) = \ell_{0-1}(z)$. On the other hand, as we assume there exists a unit vector u with margin γ over the underlying distribution, it follows that

$$\langle w_t, u\rangle = \sum_{j<t} -\ell_{\mathrm{r}}'\left(\langle w_j, z_j\rangle\right)\langle u, z_j\rangle \ge \gamma \sum_{j<t} -\ell_{\mathrm{r}}'\left(\langle w_j, z_j\rangle\right).$$

Consequently,

$$\mathbb{E}\left[\sum_{j<t}\mathcal{R}_{0-1}(w_j)\right] \le \frac{1}{\gamma}\mathbb{E}\left[\langle w_t, u\rangle\right] \le \frac{1}{\gamma}\mathbb{E}\left[\|w_t\|_2\right] \le \frac{1}{\gamma}\sqrt{\mathbb{E}\left[\|w_t\|_2^2\right]}$$
$$\le \frac{1}{\gamma}\sqrt{\mathbb{E}\left[\sum_{j<t}\mathcal{R}_{0-1}(w_j)\right]},$$

which implies

$$\mathbb{E}\left[\sum_{j<t}\mathcal{R}_{0-1}(w_j)\right] \le \frac{1}{\gamma^2},$$

and thus

$$\mathbb{E}\left[\min_{j<t}\mathcal{R}_{0-1}(w_j)\right] \le \frac{1}{\gamma^2 t}.$$

In the above analysis for perceptron, we first derive an upper bound on $\mathbb{E}\left[\|w_t\|_2^2\right]$, then obtain a lower bound on $\mathbb{E}\left[\langle w_t, u\rangle\right]$ based on the linear separability assumption, and finally combine both bounds to achieve a test error bound. Next we analyze SGD with the logistic loss using a similar perceptron-style analysis; however, care is needed to make this replacement with logistic work, as detailed below. The analysis is basically from [14].

Theorem 2.1. With the logistic loss and a constant learning rate $\eta_t = 1$, SGD ensures

$$\mathbb{E}\left[\min_{j<t}\mathcal{R}_{0-1}(w_j)\right] \le \frac{4\ln(t)}{\gamma^2 t} + \frac{4}{\gamma t}.$$

To prove Theorem 2.1, we first prove the following general result. Note that it requires $-\ell' \le \ell$, which is not true for the ReLU loss; this is the key difference between the ReLU loss and logistic loss.

Lemma 2.1. Given a convex loss ℓ with $0 \le -\ell' \le 1$ and $-\ell' \le \ell$, for any $w \in \mathbb{R}^d$ and $t \ge 1$,

$$\|w_t - w\|_2^2 \le \|w\|_2^2 + 2\sum_{j<t}\ell\left(\langle w, z_j\rangle\right).$$

Suppose furthermore $\ell(z) \le e^{-z}$, let $u_t = u\ln(t)/\gamma$,

$$\|w_t - u_t\|_2^2 \le \frac{\ln(t)^2}{\gamma^2} + 2.$$

Proof. By the SGD update rule and convexity of ℓ, for any $w \in \mathbb{R}^d$,

$$\begin{aligned}\|w_{j+1} - w\|_2^2 &= \|w_j - w\|_2^2 - 2\left\langle \ell'\left(\langle w_j, z_j\rangle\right) z_j, w_j - w\right\rangle + \ell'\left(\langle w_j, z_j\rangle\right)^2 \|z_j\|_2^2 \\ &\le \|w_j - w\|_2^2 - 2\left(\ell\left(\langle w_j, z_j\rangle\right) - \ell\left(\langle w, z_j\rangle\right)\right) + \ell'\left(\langle w_j, z_j\rangle\right)^2 \|z_j\|_2^2.\end{aligned}$$

Furthermore, since $\|z_j\|_2 \le 1$, and $0 \le -\ell' \le 1$, and $-\ell' \le \ell$,

$$\ell'\left(\langle w_j, z_j\rangle\right)^2 \|z_j\|_2^2 \le -\ell'\left(\langle w_j, z_j\rangle\right) \le \ell\left(\langle w_j, z_j\rangle\right).$$

As a result,

$$\begin{aligned}\|w_{j+1} - w\|_2^2 &\le \|w_j - w\|_2^2 - 2\left(\ell\left(\langle w_j, z_j\rangle\right) - \ell\left(\langle w, z_j\rangle\right)\right) + \ell\left(\langle w_j, z_j\rangle\right) \\ &= \|w_j - w\|_2^2 - \ell\left(\langle w_j, z_j\rangle\right) + 2\ell\left(\langle w, z_j\rangle\right) \\ &\le \|w_j - w\|_2^2 + 2\ell\left(\langle w, z_j\rangle\right).\end{aligned}$$

Take the sum from step 0 to $t-1$, the first claim of Lemma 2.1 is proved.

By the linear separability assumption, $\langle u, z_j\rangle \ge \gamma$. Therefore $\langle u_t, z_j\rangle \ge \ln(t)$, and since $\ell(z) \le e^{-z}$,

$$2\sum_{j<t} \ell\left(\langle u_t, z_j\rangle\right) \le 2\sum_{j<t} \exp\left(-\ln(t)\right) \le \frac{2}{t}\sum_{j<t} 1 \le 2.$$

QED.

Lemma 2.1 gives an upper bound on $\|w_t - u_t\|_2^2$; to finish a perceptron-style analysis, we further need a lower bound on $\langle w_t - u_t, u\rangle$.

Lemma 2.2. For all $t \ge 1$, it holds that

$$\mathbb{E}\left[\|w_t - u_t\|_2\right] \ge \mathbb{E}\left[\langle w_t - u_t, u\rangle\right] \ge \gamma\mathbb{E}\left[\sum_{j<t} \mathcal{Q}(w_j)\right] - \frac{\ln(t)}{\gamma}.$$

Proof. We have

$$\begin{aligned}\langle w_t - u_t, u\rangle &= \langle w_t, u\rangle - \langle u_t, u\rangle \\ &= \sum_{j<t} -\ell'\left(\langle w_j, z_j\rangle\right)\langle z_j, u\rangle - \frac{\ln(t)}{\gamma} \\ &\geq \gamma \sum_{j<t} -\ell'\left(\langle w_j, z_j\rangle\right) - \frac{\ln(t)}{\gamma}.\end{aligned}$$

Take the expectation on both sides, and recall that $\mathcal{Q}(w) := \mathbb{E}\left[-\ell'\left(y\langle w, x\rangle\right)\right]$ (cf. eq. (1.2)), the proof is done. QED.

Now we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. Lemmas 2.1 and 2.2 imply

$$\gamma\mathbb{E}\left[\sum_{j<t}\mathcal{Q}(w_j)\right] - \frac{\ln(t)}{\gamma} \leq \sqrt{\frac{\ln(t)^2}{\gamma^2} + 2} \leq \frac{\ln(t)}{\gamma} + 2,$$

which implies

$$\mathbb{E}\left[\sum_{j<t}\mathcal{Q}(w_j)\right] \leq \frac{2\ln(t)}{\gamma^2} + \frac{2}{\gamma},$$

and thus

$$\mathbb{E}\left[\min_{j<t}\mathcal{Q}(w_j)\right] \leq \frac{2\ln(t)}{\gamma^2 t} + \frac{2}{\gamma t}.$$

As noted in [36], $\ell_{0-1}(z) \leq -2\ell'_{\log}(z)$, and therefore

$$\mathbb{E}\left[\min_{j<t}\mathcal{R}_{0-1}(w_j)\right] \leq \frac{4\ln(t)}{\gamma^2 t} + \frac{4}{\gamma t}.$$

QED.

2.2 A PRIMAL-DUAL ANALYSIS OF THE IMPLICIT BIAS

In the previous section, we showed an $\widetilde{O}(1/t)$ test error bound for SGD with the logistic loss. It turns out that a similar test error bound can also be showed for GD [37]. On the

other hand, in addition to a test error bound, it is possible to give a finer characterization of the GD iterates: it was showed in [16] that with an exponentially-tailed loss such as the exponential loss or logistic loss, it holds that

$$\lim_{t\to\infty} \|w_t\|_2 = \infty, \quad \text{and} \quad \lim_{t\to\infty} \frac{w_t}{\|w_t\|_2} = u^*.$$

Here we give an alternative proof of this result using a primal-dual framework. More precisely, we will show that for each GD iterate w_t, we can define a corresponding *dual* iterate q_t, and the update on q_t is a *mirror descent (multiplicative weight)* update on a certain dual objective whose optimum is given exactly by the maximum margin. This observation gives a nice intuition why GD converges to the maximum-margin solution, and also allows us to design fast margin maximization algorithms (cf. Theorems 2.3 and 2.5) and prove an alignment property for deep homogeneous networks (cf. Section 5.2).

2.2.1 A primal-dual framework

For simplicity, here we focus on the exponential loss $\ell_{\exp}$; however, as discussed in [19], the analysis can also be extended to the logistic loss. For simplicity, let $z_i := y_i x_i$, and collect them into a matrix $Z \in \mathbb{R}^{n\times d}$. Moreover, given $\xi \in \mathbb{R}^n$, let $\psi(\xi) := \ln\left(\sum_{i=1}^n \exp(\xi_i)\right)$ denote the ln-sum-exp function. Given a GD iterate w_t, let $p_t := -Zw_t$, and we further define a dual variable $q_t \in \Delta_n$ (the probability simplex) by

$$q_{t,i} := \frac{\exp\left(-\langle w_t, z_i\rangle\right)}{\sum_{i'=1}^n \exp\left(-\langle w_t, z_{i'}\rangle\right)}, \quad \text{or equivalently} \quad q_t := \nabla\psi(p_t).$$

It then holds that

$$w_{t+1} = w_t - \eta_t \nabla\widehat{\mathcal{R}}(w_t) = w_t + \eta_t \sum_{i=1}^n \frac{1}{n} \exp\left(-\langle w_t, z_i\rangle\right) z_i = w_t + \hat{\eta}_t Z^\top q_t,$$

where we let $\hat{\eta}_t := \eta_t \widehat{\mathcal{R}}(w_t)$. This further implies that

$$p_{t+1} = p_t - Z(w_{t+1} - w_t) = p_t - \hat{\eta}_t Z Z^\top q_t = p_t - \hat{\eta}_t \nabla\phi(q_t),$$

where $\phi(q) := \|Z^\top q\|_2^2/2$. To sum up, we have

$$p_{t+1} = p_t - \hat{\eta}_t \nabla\phi(q_t), \quad \text{and} \quad q_{t+1} := \nabla\psi(p_{t+1}).$$

This update on q_t is exactly a mirror descent or dual averaging [38] update on the dual objective ϕ, with learning rate $\hat{\eta}_t$ and an entropy regularizer. In the following, we will see that ϕ is the key potential function in our primal-dual framework; moreover, it can be generalized to the nonlinear case (cf. eq. (5.12)), which will be used in our implicit bias analysis for deep homogeneous networks.

First, we note that the optimum of ϕ is exactly characterized by the maximum margin.

Lemma 2.3. It holds that

$$\gamma^* := \max_{\|u\|_2 \le 1} \min_{1 \le i \le n} (Zu)_i = \min_{q \in \Delta_n} \left\| Z^\top q \right\|_2,$$

and there exists a unique primal optimal solution u^*, such that for any dual optimal solution q^*, it holds that $Z^\top q^* = \gamma^* u^*$.

Proof. Given a convex set C, recall that ι_C denote the indicator function, i.e., $\iota_C(x) = 0$ if $x \in C$, and $\iota_C(x) = \infty$ if $x \notin C$. We note the following convex conjugate pairs:

$$\begin{aligned} \iota^*_{\Delta_n}(v) &= \sup_{u \in \Delta_n} \langle v, u \rangle = \max_{1 \le i \le n} v_i, \\ (\|\cdot\|_2)^*(q) &= \iota_{\|\cdot\|_2 \le 1}(q). \end{aligned}$$

This gives the Fenchel strong duality [39, Theorem 3.3.5]

$$\begin{aligned} \min \left(\|Z^\top q\|_2 + \iota_{\Delta_n}(q) \right) &= \max -\iota_{\|\cdot\|_2 \le 1}(-w) - \iota^*_{\Delta_n}(Zw) \\ &= \max \left\{ -\max_i (Zw)_i : \|w\|_2 \le 1 \right\} \\ &= \max \left\{ \min_i (Z(-w))_i : \|w\|_2 \le 1 \right\} \\ &= \max \left\{ \min_i (Zu)_i : \|u\|_2 \le 1 \right\} \end{aligned}$$

Now consider an arbitrary optimal primal-dual pair (w^*, q^*). Fenchel-Young's inequality [39, Proposition 3.3.4] implies for any w and q that

$$\|Z^\top q\|_2 + \iota_{\|\cdot\|_2 \le 1}(-w) + \iota_{\Delta_n}(q) + \iota^*_{\Delta_n}(Zw) \ge \left\langle Z^\top q, -w \right\rangle + \langle q, Zw \rangle = 0.$$

The optimal pair (w^*, q^*) satisfies the above inequality with an equality, and it follows from [39, Proposition 3.3.4] that $Z^\top q^* \in \partial \left(\iota_{\|\cdot\|_2 \le 1} \right)(-w^*)$, meaning $Z^\top q^*$ and $-w^*$ have the same direction. Since $\left\| Z^\top q^* \right\|_2 = \gamma^*$, and let $u^* = -w^*$, we have $Z^\top q^* = \gamma^* u^*$. Since the above

argument holds for any optimal primal-dual pair, it follows that u^* is unique. QED.

As mentioned above, since the update on q_t is a mirror descent update which tries to minimize ϕ, and it follows from Lemma 2.3 that the optimum of ϕ is exactly characterized by (u^*, q^*), this would give an intuitive explanation of the implicit bias phenomenon if the dual objective ϕ can be globally minimized by q_t. Next we show that this is indeed the case, along with a primal-dual convergence rate, based on [19, Theorem 1]. It will also be the basis of our fast $O(1/t)$ margin maximization rate for normalized GD (cf. Theorem 2.3).

Let ψ^* denote the convex conjugate of ψ; formally, ψ^* denotes the entropy function, where given $q \in \Delta_n$, we have $\psi^*(q) := \sum_{i=1}^n q_i \ln(q_i)$. Given $q \in \Delta_n$, the Bregman distance between q and q_t is defined as

$$D_{\psi^*}(q, q_t) := \psi^*(q) - \psi^*(q_t) - \langle p_t, q - q_t \rangle.$$

Since ψ^* denotes the entropy function, D_{ψ^*} is actually just the KL-divergence. Here is our main convergence result.

Theorem 2.2. For all $q \in \Delta_n$, if $\hat{\eta}_t \leq 1$, then the following results hold:

1. Dual convergence: for all $t \geq 0$,

$$\phi(q_{t+1}) \leq \phi(q_t), \quad \text{and} \quad \hat{\eta}_t \left(\phi(q_{t+1}) - \phi(q)\right) \leq D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}).$$

   As a result, for all $t > 0$,

$$\phi(q_t) - \phi(q) \leq \frac{D_{\psi^*}(q, q_0) - D_{\psi^*}(q, q_t)}{\sum_{j<t} \hat{\eta}_j} \leq \frac{D_{\psi^*}(q, q_0)}{\sum_{j<t} \hat{\eta}_j}.$$

2. Primal convergence: for all $t \geq 0$,

$$\psi(p_t) - \psi(p_{t+1}) \geq \hat{\eta}_t \left(\phi(q_t) + \phi(q_{t+1})\right) = \frac{\hat{\eta}_t}{2} \left\| Z^\top q_t \right\|_2^2 + \frac{\hat{\eta}_t}{2} \left\| Z^\top q_{t+1} \right\|_2^2,$$

   and thus if $\hat{\eta}_t$ is nonincreasing, then

$$\psi(p_0) - \psi(p_t) \geq \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\|_2^2 - \frac{\hat{\eta}_0}{2} \left\| Z^\top q_0 \right\|_2^2 + \frac{\hat{\eta}_t}{2} \left\| Z^\top q_t \right\|_2^2.$$

   This rate is tight up to a constant, since $\psi(p_0) - \psi(p_t) \leq \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\|_2^2$.

Here are some comments on Theorem 2.2.

- If we let $\hat{\eta}_t = 1$, then we get an $O(1/t)$ dual convergence rate. By contrast, [40] considered boosting, and can only handle step size $\hat{\eta}_t \propto 1/\sqrt{t+1}$ and give an $\widetilde{O}(1/\sqrt{t})$ dual rate. This is because the dual objective $\phi(q) := \left\|Z^\top q\right\|_2^2/2$ for gradient descent is smooth, while for boosting the dual objective is given by $\left\|Z^\top q\right\|_\infty^2/2$, which is non-smooth. In some sense, we can handle a constant $\hat{\eta}_t$ and prove a faster rate because *both the primal objective ψ and the dual objective ϕ are smooth.*

- Moreover, the primal and dual smoothness allow us to prove a super tight primal convergence rate for ψ. By contrast, if we use a standard smoothness guarantee, then the error term (compared with the upper bound on $\psi(p_0) - \psi(p_t)$) can be as large as $\sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|_2^2/2$ (cf. Lemma 2.5). While a constant factor does not hurt the risk bound too much, it can stop us from proving an $O(1/t)$ margin maximization rate (cf. Section 2.2.2).

- For the exponential loss (and other exponentially-tailed losses), [16] proved that w_t converges to the maximum margin direction. This is called an "implicit bias" result since it does not follow from classical results such as risk minimization, and requires a nontrivial proof tailored to the exponential function. By contrast, Theorem 2.2 explicitly shows that the dual iterates minimize the dual objective ϕ, and the minimum of ϕ is given exactly by the maximum margin as showed by Lemma 2.3, which gives an intuitive explanation of the implicit bias phenomenon.

Next we prove Theorem 2.2. One of the key properties we use is the ℓ_1 smoothness of ϕ.

Lemma 2.4. The function $\phi : \mathbb{R}^n \to \mathbb{R}$ given by $\phi(\theta) := \left\|Z^\top \theta\right\|_2^2/2$ is 1-smooth with respect to the ℓ_1 norm.

Proof. For any $\theta, \theta' \in \mathbb{R}^n$, using the Cauchy-Schwarz inequality and $\|z_i\| \le 1$,

$$\begin{aligned}\left\|\nabla\phi(\theta) - \nabla\phi(\theta')\right\|_\infty = \left\|ZZ^\top(\theta - \theta')\right\|_\infty &= \max_{1\le i\le n}\left|\left\langle Z^\top(\theta-\theta'), z_i\right\rangle\right| \\ &\le \max_{1\le i\le n}\left\|Z^\top(\theta-\theta')\right\|_2 \|z_i\|_2 \\ &\le \left\|Z^\top(\theta-\theta')\right\|_2.\end{aligned}$$

Furthermore, by the triangle inequality and $\|z_i\|_2 \le 1$,

$$\left\|Z^\top(\theta-\theta')\right\|_2 \le \sum_{i=1}^n \left|\theta_i - \theta_i'\right| \|z_i\|_2 \le \sum_{i=1}^n \left|\theta_i - \theta_i'\right| = \left\|\theta - \theta'\right\|_1.$$

Therefore ϕ is 1-smooth with respect to the ℓ_1 norm. QED.

Next, here are some standard results we need, from the smoothness of ψ.

Lemma 2.5. We have

$$\psi(p_{t+1}) - \psi(p_t) \le -\hat{\eta}_t \left\| Z^\top q_t \right\|_2^2 + \frac{\hat{\eta}_t^2}{2} \left\| Z^\top q_t \right\|_2^2 \qquad \text{and} \qquad D_{\psi^*}(q_{t+1}, q_t) \ge \frac{1}{2} \|q_{t+1} - q_t\|_1^2.$$

Proof. We first claim that ψ is 1-smooth with respect to the ℓ_∞ norm. Similarly to the proof of [41, Lemma 14], it is enough to show that for any $\xi, v \in \mathbb{R}^n$, it holds that $v^\top \nabla^2 \psi(\xi) v \le \|v\|_\infty^2$. Note that

$$\nabla^2 \psi(\xi) = \operatorname{diag}\left(\nabla \psi(\xi)\right) - \nabla \psi(\xi) \nabla \psi(\xi)^\top,$$

where

$$\nabla \psi(\xi) = \left(\frac{e^{\xi_1}}{\sum_{i=1}^n e^{\xi_i}}, \dots, \frac{e^{\xi_n}}{\sum_{i=1}^n e^{\xi_i}} \right).$$

Therefore it is enough to show that

$$\sum_{i=1}^n \nabla \psi(\xi)_i v_i^2 \le \max_{1 \le i \le n} v_i^2,$$

which is true since $\sum_{i=1}^n \nabla \psi(\xi)_i = 1$.

Next we prove the main claims. Since ψ is 1-smooth with respect to the ℓ_∞ norm,

$$\begin{aligned}
\psi(p_{t+1}) - \psi(p_t) &\le \left\langle \nabla \psi(p_t), p_{t+1} - p_t \right\rangle + \frac{1}{2} \|p_{t+1} - p_t\|_\infty^2 \\
&= \left\langle q_t, -\hat{\eta}_t Z Z^\top q_t \right\rangle + \frac{\hat{\eta}_t^2}{2} \left\| Z Z^\top q_t \right\|_\infty^2 \\
&= -\hat{\eta}_t \left\| Z^\top q_t \right\|_2^2 + \frac{\hat{\eta}_t^2}{2} \left\| Z Z^\top q_t \right\|_\infty^2.
\end{aligned}$$

Moreover, since $\|z_i\|_2 \le 1$,

$$\left\| Z Z^\top q_t \right\|_\infty = \max_{1 \le i \le n} \left| \left\langle Z^\top q_t, z_i \right\rangle \right| \le \max_{1 \le i \le n} \left\| Z^\top q_t \right\|_2 \|z_i\|_2 \le \left\| Z^\top q_t \right\|_2.$$

As a result,

$$\psi(p_{t+1}) - \psi(p_t) \le -\hat{\eta}_t \left\| Z^\top q_t \right\|_2^2 + \frac{\hat{\eta}_t^2}{2} \left\| Z^\top q_t \right\|_2^2.$$

On the second claim, note that since ψ is 1-smooth with respect to the ℓ_∞ norm, [42, Lemma 2.19] implies that ψ^* is 1-strongly convex with respect to the ℓ_1 norm, and in particular $D_{\psi^*}(q_{t+1}, q_t) \geq \|q_{t+1} - q_t\|_1^2/2$. QED.

Next is a standard result for mirror descent.

Lemma 2.6. For any $t \geq 0$ and $q \in \Delta_n$, it holds that

$$\hat{\eta}_t \left(\phi(q_t) - \phi(q)\right) \leq \left\langle \hat{\eta}_t \nabla\phi(q_t), q_t - q_{t+1} \right\rangle - D_{\psi^*}(q_{t+1}, q_t) + D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}).$$

Moreover, q_{t+1} is the unique minimizer of

$$h(q) := \phi(q_t) + \left\langle \nabla\phi(q_t), q - q_t \right\rangle + \frac{1}{\hat{\eta}_t} D_{\psi^*}(q, q_t),$$

and specifically $h(q_{t+1}) \leq h(q_t) = \phi(q_t)$.

Proof. Since ϕ is convex, we have

$$\hat{\eta}_t \left(\phi(q_t) - \phi(q)\right) \leq \left\langle \hat{\eta}_t \nabla\phi(q_t), q_t - q \right\rangle = \left\langle \hat{\eta}_t \nabla\phi(q_t), q_t - q_{t+1} \right\rangle + \left\langle \hat{\eta}_t \nabla\phi(q_t), q_{t+1} - q \right\rangle.$$

Recall that $p_{t+1} = p_t - \hat{\eta}_t Z Z^\top q_t = p_t - \hat{\eta}_t \nabla\phi(q_t)$, therefore

$$\begin{aligned} \hat{\eta}_t \left(\phi(q_t) - \phi(q)\right) &\leq \left\langle \hat{\eta}_t \nabla\phi(q_t), q_t - q_{t+1} \right\rangle + \left\langle \hat{\eta}_t \nabla\phi(q_t), q_{t+1} - q \right\rangle \\ &= \left\langle \hat{\eta}_t \nabla\phi(q_t), q_t - q_{t+1} \right\rangle + \left\langle p_t - p_{t+1}, q_{t+1} - q \right\rangle. \end{aligned}$$

It can be verified by direct expansion that

$$\left\langle p_t - p_{t+1}, q_{t+1} - q \right\rangle = D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}) - D_{\psi^*}(q_{t+1}, q_t),$$

and thus

$$\hat{\eta}_t \left(\phi(q_t) - \phi(q)\right) \leq \left\langle \hat{\eta}_t \nabla\phi(q_t), q_t - q_{t+1} \right\rangle + D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}) - D_{\psi^*}(q_{t+1}, q_t).$$

On the other claim, let ∂ denote subdifferential. We have

$$\partial h(q) = \left\{\nabla\phi(q_t)\right\} + \frac{1}{\hat{\eta}_t} \left(\partial\psi^*(q) - \{p_t\}\right).$$

Note that $q' \in \arg\min h(q)$ if and only if $0 \in \partial h(q')$, which is equivalent to $p_t - \hat{\eta}_t \nabla\phi(q_t) = p_{t+1} \in \partial\psi^*(q)$. By [43, Theorem 23.5], $p_{t+1} \in \partial\psi^*(q)$ if and only if $q = \nabla\psi(p_{t+1})$; in other words, q_{t+1} is the unique minimizer of h, and specifically $h(q_{t+1}) \leq h(q_t) = \phi(q_t)$. QED.

With Lemmas 2.4 to 2.6, we can prove Theorem 2.2.

Proof of Theorem 2.2. Since ϕ is 1-smooth with respect to the ℓ_1 norm,

$$\phi(q_{t+1}) - \phi(q_t) \leq \left\langle \nabla\phi(q_t), q_{t+1} - q_t \right\rangle + \frac{1}{2}\|q_{t+1} - q_t\|_1^2.$$

Further invoking Lemma 2.5, and that $\hat{\eta}_t \leq 1$, and the function h defined in Lemma 2.6, it follows that

$$\begin{aligned}
\phi(q_{t+1}) &\leq \phi(q_t) + \left\langle \nabla\phi(q_t), q_{t+1} - q_t \right\rangle + \frac{1}{2}\|q_{t+1} - q_t\|_1^2 \\
&\leq \phi(q_t) + \left\langle \nabla\phi(q_t), q_{t+1} - q_t \right\rangle + D_{\psi^*}(q_{t+1}, q_t) \\
&\leq \phi(q_t) + \left\langle \nabla\phi(q_t), q_{t+1} - q_t \right\rangle + \frac{1}{\hat{\eta}_t} D_{\psi^*}(q_{t+1}, q_t) \qquad (2.1) \\
&= h(q_{t+1}) \leq \phi(q_t),
\end{aligned}$$

which proves that $\phi(q_t)$ is nonincreasing.

To prove the iteration guarantee for ϕ, note that rearranging the terms of eq. (2.1) gives the following inequality

$$\hat{\eta}_t \left\langle \nabla\phi(q_t), q_{t+1} - q_t \right\rangle + D_{\psi^*}(q_{t+1}, q_t) \geq \hat{\eta}_t \left(\phi(q_{t+1}) - \phi(q_t)\right).$$

Lemma 2.6 then implies

$$\hat{\eta}_t \left(\phi(q_t) - \phi(q)\right) \leq \hat{\eta}_t \left(\phi(q_t) - \phi(q_{t+1})\right) + D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}).$$

Rearranging terms gives

$$\hat{\eta}_t \left(\phi(q_{t+1}) - \phi(q)\right) \leq D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}). \qquad (2.2)$$

Taking the sum of eq. (2.2) from 0 to $t-1$, and noting that $\phi(q_{j+1}) \geq \phi(q_t)$ for all $j < t$ since f is nonincreasing, the proof is done.

To prove the iteration guarantee for ψ, note that

$$\begin{aligned}
D_{\psi^*}(q_{t+1}, q_t) &= \psi^*(q_{t+1}) - \psi^*(q_t) - \langle p_t, q_{t+1} - q_t \rangle \\
&= \langle p_{t+1}, q_{t+1} \rangle - \psi(p_{t+1}) - \langle p_t, q_t \rangle + \psi(p_t) - \langle p_t, q_{t+1} - q_t \rangle \\
&= \psi(p_t) - \psi(p_{t+1}) - \langle q_{t+1}, p_t - p_{t+1} \rangle \\
&= \psi(p_t) - \psi(p_{t+1}) - \hat{\eta}_t \left\langle Z^\top q_t, Z^\top q_{t+1} \right\rangle.
\end{aligned}$$

Further invoking eq. (2.1), we have

$$\begin{aligned}\psi(p_t) - \psi(p_{t+1}) &\geq \hat{\eta}_t \left(\phi(q_{t+1}) - \phi(q_t) - \left\langle \nabla\phi(q_t), q_{t+1} - q_t \right\rangle + \left\langle Z^\top q_t, Z^\top q_{t+1} \right\rangle \right) \\ &= \frac{\hat{\eta}_t}{2} \left\| Z^\top q_t \right\|_2^2 + \frac{\hat{\eta}_t}{2} \left\| Z^\top q_{t+1} \right\|_2^2.\end{aligned}$$

Telescoping gives the lower bound on $\psi(p_0) - \psi(p_t)$. For the upper bound, note that ψ is convex, and thus

$$\psi(p_t) - \psi(p_{t+1}) \leq \langle q_t, p_t - p_{t+1} \rangle = \left\langle q_t, \hat{\eta}_t Z Z^\top q_t \right\rangle = \hat{\eta}_t \left\| Z^\top q_t \right\|_2^2.$$

QED.

2.2.2 $O(1/t)$ margin maximization rate for normalized GD

Here we show an $O(1/t)$ margin maximization rate for GD with $\hat{\eta}_t = 1$ (equivalently $\eta_t = 1/\widehat{\mathcal{R}}(w_t)$), based on [19, Theorem 7]. We let $w_0 := 0$, though our analysis can be easily extended to handle nonzero initialization.

Theorem 2.3. If $\hat{\eta}_t = \eta_t \widehat{\mathcal{R}}(w_t) \leq 1$ is nonincreasing and $w_0 = 0$, then

$$\frac{\min_{1 \leq i \leq n} y_i \langle w_t, x_i \rangle}{\|w_t\|_2} \geq \frac{-\psi(-Z w_t)}{\|w_t\|_2} \geq \gamma - \frac{\ln(n) + 1}{\gamma \sum_{j<t} \hat{\eta}_j}.$$

Margin maximization has been analyzed in many settings: [18] proved that for any $\epsilon > 0$, the margin can be maximized by coordinate descent to $\gamma - \epsilon$ with an $O(1/t)$ rate, while [44] showed an $\widetilde{O}(1/\sqrt{t})$ margin maximization rate for gradient descent by letting $\hat{\eta}_t \propto 1/\sqrt{t+1}$. They also analyzed the quantity $-\psi(-Zw_t)/\|w_t\|_2$, but used Lemma 2.5. If we let $\hat{\eta}_t$ be a constant in Lemma 2.5, the error term $\sum_{j<t} \frac{\beta \hat{\eta}_j^2}{2} \left\| Z^\top q_j \right\|_2^2$ will be too large to prove exact margin maximization, while if we let $\hat{\eta}_t = 1/\sqrt{t+1}$, then the error term is $O\left(\ln(t)\right)$, but only an $O\left(\ln(t)/\sqrt{t}\right)$ rate can be obtained. By contrast, the proof of Theorem 2.3 uses the tighter guarantee given by Theorem 2.2, which always has a bounded error term.

We can also directly run (sub)gradient descent on the (negative) margin function over the unit ℓ_2 ball:

$$\min_{\|w\|_2 \leq 1} \max_{1 \leq i \leq n} y_i \langle w, x_i \rangle.$$

This gives the batch perceptron algorithm [21], which has $O(1/\sqrt{t})$ convergence rate. [45]

gives a normalized Perceptron algorithm, which also has $O(1/\sqrt{t})$ convergence rate. Finally, we may also apply the ellipsoid method to the margin function; to ensure an ϵ additive error, it requires $O(d^2 \ln(1/\epsilon))$ steps [9, Theorem 2.4]. In the case where d is large, our convergence rate in Theorem 2.3 is still faster.

Proof of Theorem 2.3. Theorem 2.2 and lemma 2.3 imply that

$$\begin{aligned}
-\psi(p_t) &\geq -\psi(p_0) + \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\|_2^2 - \frac{\hat{\eta}_0}{2} \left\| Z^\top q_0 \right\|_2^2 \\
&\geq -\psi(p_0) + \gamma \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\|_2 - \frac{\hat{\eta}_0}{2} \left\| Z^\top q_0 \right\|_2^2 .
\end{aligned}$$

Then we have

$$\begin{aligned}
\frac{-\psi(-Zw_t)}{\|w_t\|_2} = \frac{-\psi(p_t)}{\|w_t\|_2} &\geq \frac{-\psi(p_0) + \gamma \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\|_2 - \frac{\hat{\eta}_0}{2} \left\| Z^\top q_0 \right\|_2^2}{\|w_t\|_2} \\
&= \gamma \cdot \frac{\sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\|_2}{\|w_t\|_2} - \frac{\psi(p_0) + \frac{\hat{\eta}_0}{2} \left\| Z^\top q_0 \right\|_2^2}{\|w_t\|_2}.
\end{aligned}$$

It follows from the triangle inequality that $\|w_t\|_2 \leq \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\|_2$. Moreover, $\psi(p_0) = \ln(n)$, and $\left\| Z^\top q_0 \right\|_2 \leq 1$ since $\|z_i\|_2 \leq 1$. Therefore we have

$$\frac{-\psi(-Zw_t)}{\|w_t\|_2} \geq \gamma - \frac{\ln(n) + 1}{\|w_t\|_2}.$$

Furthermore, note that $\|w_t\|_2 \geq \langle w_t, u^* \rangle$, and

$$\langle w_{j+1} - w_j, u^* \rangle = \hat{\eta}_j \left\langle Z^\top q_j, u^* \right\rangle = \hat{\eta}_j \langle Z u^*, q_j \rangle \geq \hat{\eta}_j \gamma,$$

which implies

$$\frac{-\psi(-Zw_t)}{\|w_t\|_2} \geq \gamma - \frac{\ln(n) + 1}{\gamma \sum_{j<t} \hat{\eta}_j}.$$

Finally, note that for the exponential loss,

$$\psi(-Zw) = \ln \left(\sum_{i=1}^n \exp\left(\langle -z_i, w \rangle \right) \right) \geq \ln \left(\exp \left(\max_{1 \leq i \leq n} \langle -z_i, w \rangle \right) \right) = - \min_{1 \leq i \leq n} \langle z_i, w \rangle,$$

and thus $\min_{1 \leq i \leq n} \langle z_i, w \rangle = \min_{1 \leq i \leq n} y_i \langle w, x_i \rangle \geq -\psi(-Zw)$. QED.

2.3 FAST MARGIN MAXIMIZATION VIA DUAL ACCELERATION

In this section, we further exploit the primal-dual framework developed in the previous section, and designed a momentum-style algorithm that achieves an $\widetilde{O}(1/t^2)$ margin maximization rate. This section is based on [20].

The algorithm is given by

$$g_t := \beta_t \left(g_{t-1} + \frac{\nabla \widehat{\mathcal{R}}(w_t)}{\widehat{\mathcal{R}}(w_t)} \right), \quad \text{and} \quad w_{t+1} := w_t - \theta_t \left(g_t + \frac{\nabla \widehat{\mathcal{R}}(w_t)}{\widehat{\mathcal{R}}(w_t)} \right). \tag{2.3}$$

Our main result is that these iterates, with a proper choice of θ_t and β_t, can maximize the margin at a rate of $\widetilde{\mathcal{O}}(1/t^2)$, whereas prior work had a rate of $\mathcal{O}(1/t)$ at best. The key idea is to reverse the primal-dual relationship discussed in the previous section: therefore we show that primal normalized GD is equivalent to dual mirror descent, but here we start from the dual, and apply Nesterov acceleration to make dual optimization faster, and then translate the dual iterates into the momentum form in eq. (2.3). Note that if our goal is just to accelerate dual optimization, then it is natural to apply Nesterov's method; however, here our goal is to accelerate (primal) margin maximization – it was unclear whether the momentum method changes the implicit bias, and our margin analysis is very different from the standard analysis of Nesterov's method.

2.3.1 A Unified Analysis of Normal and Accelerated Dual Averaging

We first present some general results that will be used throughout our analysis. Consider a convex function ϕ (which can be arbitrary), and a convex set C, such that ϕ is well-defined defined and 1-smooth with respect to norm $\|\cdot\|$ on C. Moreover, suppose $\omega : C \to \mathbb{R}$ is differentiable, closed, proper, and 1-strongly convex with respect to the same norm $\|\cdot\|$. We maintain three sequences q_t, μ_t, ν_t: initialize $\mu_0 = q_0 \in C$, and for $t \geq 0$, let

$$\begin{aligned} \nu_t &:= (1-\lambda_t)\mu_t + \lambda_t q_t, \\ q_{t+1} &:= \operatorname*{arg\,min}_{q \in C} \left(\phi(q_t) + \left\langle \nabla \phi(\nu_t), q - q_t \right\rangle + \frac{\lambda_t}{\theta_t} D_\omega(q, q_t) \right), \\ \mu_{t+1} &:= (1-\lambda_t)\mu_t + \lambda_t q_{t+1}, \end{aligned} \tag{2.4}$$

where $\lambda_t, \theta_t \in (0, 1]$, and $D_\omega(q, q') := \omega(q) - \omega(q') - \left\langle \nabla \omega(q'), q - q' \right\rangle$ denotes the Bregman distance.

The above update to q_t resembles the mirror descent update. We can instead use a

dual-averaging update, which does not require differentiability of ω: first note that since ω is strongly convex, its convex conjugate ω^* is smooth [42, lemma 2.19], and thus it is well-defined and differentiable on the whole Euclidean space. For any initialization p_0, let $q_0 := \nabla\omega^*(p_0)$ and for $t \geq 0$, let

$$p_{t+1} := p_t - \frac{\theta_t}{\lambda_t}\nabla\phi(\nu_t), \quad \text{and} \quad q_{t+1} := \nabla\omega^*(p_{t+1}). \tag{2.5}$$

Note that it is exactly the original update to q_t if for all $t \geq 0$ and $q \in C$, we define $D_\omega(q, q_t) := \omega(q) - \omega(q_t) - \langle p_t, q - q_t\rangle$. Below we will analyze this dual-averaging update.

The following result is crucial to our analysis. It combines a dual averaging analysis and a Nesterov analysis: when $\theta_t = 1$, it is basically [34, eq. (24)], and choosing a proper λ_t would give us acceleration; on the other hand, we further handle the case of $\lambda_t = 1$, when it becomes the usual convergence result for dual averaging.

Lemma 2.7. If $\lambda_t, \theta_t \in (0, 1]$ for all $t \geq 0$, then for all $t \geq 1$ and $q \in C$,

$$\begin{aligned}
&\frac{\theta_{t-1}}{\lambda_{t-1}^2}\left(\phi(\mu_t) - \phi(q)\right) + \sum_{j=1}^{t-1}\left(\frac{\theta_{j-1}}{\lambda_{j-1}^2} - \frac{\theta_j(1-\lambda_j)}{\lambda_j^2}\right)\left(\phi(\mu_j) - \phi(q)\right) \\
\leq\ & D_\omega(q, q_0) - D_\omega(q, q_t) + \frac{\theta_0(1-\lambda_0)}{\lambda_0^2}\left(\phi(\mu_0) - \phi(q)\right).
\end{aligned}$$

To prove Lemma 2.7, we first recall the following standard result on mirror descent. Its proof includes direct expansion and calculation, and thus is omitted.

Lemma 2.8. For all $t \geq 0$ and $q \in C$,

$$\langle p_t - p_{t+1}, q_{t+1} - q\rangle = D_\omega(q, q_t) - D_\omega(q, q_{t+1}) - D_\omega(q_{t+1}, q_t).$$

Now we are ready to prove Lemma 2.7.

Proof of Lemma 2.7. For any $t \geq 0$ and $q \in C$,

$$\begin{aligned}
\phi(\nu_t) - \phi(q) &\leq \left\langle \nabla\phi(\nu_t), \nu_t - q\right\rangle \\
&= \left\langle \nabla\phi(\nu_t), \nu_t - q_t\right\rangle + \left\langle \nabla\phi(\nu_t), q_t - q\right\rangle \\
&= \frac{1-\lambda_t}{\lambda_t}\left\langle \nabla\phi(\nu_t), \mu_t - \nu_t\right\rangle + \left\langle \nabla\phi(\nu_t), q_t - q\right\rangle \\
&\leq \frac{1-\lambda_t}{\lambda_t}\left(\phi(\mu_t) - \phi(\nu_t)\right) + \left\langle \nabla\phi(\nu_t), q_t - q\right\rangle.
\end{aligned} \tag{2.6}$$

Moreover,

$$\begin{aligned}\left\langle \nabla\phi(\nu_t), q_t - q\right\rangle &= \left\langle \nabla\phi(\nu_t), q_t - q_{t+1}\right\rangle + \left\langle \nabla\phi(\nu_t), q_{t+1} - q\right\rangle \\ &= \left\langle \nabla\phi(\nu_t), q_t - q_{t+1}\right\rangle + \frac{\lambda_t}{\theta_t}\langle p_t - p_{t+1}, q_{t+1} - q_t\rangle \\ &= \left\langle \nabla\phi(\nu_t), q_t - q_{t+1}\right\rangle - \frac{\lambda_t}{\theta_t}D_\omega(q_{t+1}, q_t) + \frac{\lambda_t}{\theta_t}\left(D_\omega(q, q_t) - D_\omega(q, q_{t+1})\right),\end{aligned} \tag{2.7}$$

where we use Lemma 2.8 in the last step. Next by 1-smoothness of ϕ and 1-strong convexity of ω, we have

$$\begin{aligned}\phi(\mu_{t+1}) - \phi(\nu_t) &\le \left\langle \nabla\phi(\nu_t), \mu_{t+1} - \nu_t\right\rangle + \frac{1}{2}\|\mu_{t+1} - \nu_t\|^2 \\ &= \lambda_t\left\langle \nabla\phi(\nu_t), q_{t+1} - q_t\right\rangle + \frac{\lambda_t^2}{2}\|q_{t+1} - q_t\|^2 \\ &\le \lambda_t\left\langle \nabla\phi(\nu_t), q_{t+1} - q_t\right\rangle + \frac{\lambda_t^2}{2\theta_t}\|q_{t+1} - q_t\|^2 \\ &\le \lambda_t\left\langle \nabla\phi(\nu_t), q_{t+1} - q_t\right\rangle + \frac{\lambda_t^2}{\theta_t}D_\omega(q_{t+1}, q_t),\end{aligned}$$

and therefore

$$\left\langle \nabla\phi(\nu_t), q_t - q_{t+1}\right\rangle - \frac{\lambda_t}{\theta_t}D_\omega(q_{t+1}, q_t) \le \frac{1}{\lambda_t}\left(\phi(\nu_t) - \phi(\mu_{t+1})\right). \tag{2.8}$$

Then eqs. (2.6) to (2.8) imply

$$\begin{aligned}\phi(\nu_t) - \phi(q) &\le \left\langle \nabla\phi(\nu_t), \nu_t - q\right\rangle \\ &\le \frac{1-\lambda_t}{\lambda_t}\left(\phi(\mu_t) - \phi(\nu_t)\right) + \frac{1}{\lambda_t}\left(\phi(\nu_t) - \phi(\mu_{t+1})\right) + \frac{\lambda_t}{\theta_t}\left(D_\omega(q, q_t) - D_\omega(q, q_{t+1})\right),\end{aligned} \tag{2.9}$$

and rearranging terms gives

$$\frac{1}{\lambda_t}\left(\phi(\mu_{t+1}) - \phi(q)\right) - \frac{1-\lambda_t}{\lambda_t}\left(\phi(\mu_t) - \phi(q)\right) \le \frac{\lambda_t}{\theta_t}\left(D_\omega(q, q_t) - D_\omega(q, q_{t+1})\right).$$

Multiply both sides by θ_t/λ_t, and then take the sum from step 0 to $t-1$, the proof is finished. QED.

Next we invoke Lemma 2.7 to get concrete rates. We further make the following require-

Algorithm 2.1: Momentum-style margin maximization.

Input: data matrix $Z \in \mathbb{R}^{n\times d}$, step size $(\theta_t)_{t=0}^{\infty}$, momentum factor $(\beta_t)_{t=0}^{\infty}$.
Initialize: $w_0 = g_{-1} = (0,\dots,0) \in \mathbb{R}^d$, $q_0 = (\frac{1}{n},\dots,\frac{1}{n}) \in \Delta_n$.
for $t = 0, 1, 2, \dots$ **do**
  $g_t \leftarrow \beta_t(g_{t-1} - Z^\top q_t)$.
  $w_{t+1} \leftarrow w_t - \theta_t\left(g_t - Z^\top q_t\right)$.
  $q_{t+1} \propto \exp(-Zw_{t+1})$, and $q_{t+1} \in \Delta_n$.
end for

ment on λ_t:

$$\lambda_0 := 1, \quad \text{and} \quad \frac{1}{\lambda_t^2} - \frac{1}{\lambda_t} \le \frac{1}{\lambda_{t-1}^2} \text{ for all } t \ge 1. \tag{2.10}$$

Note that by this construction,

$$\frac{1}{\lambda_t^2} \le \frac{1}{\lambda_0^2} + \sum_{j=1}^{t} \frac{1}{\lambda_t} = \sum_{j=0}^{t} \frac{1}{\lambda_t}. \tag{2.11}$$

Theorem 2.4. With eq. (2.10) satisfied and $\theta_t = 1$, for all $t \ge 1$ and $q^* \in \arg\min_{q\in C} \phi(q)$,

$$\phi(\mu_t) - \phi(q^*) \le \lambda_{t-1}^2 \left(D_\omega(q^*, q_0) - D_\omega(q^*, q_t)\right) \le \lambda_{t-1}^2 D_\omega(q^*, q_0).$$

In particular, if $\lambda_t = 2/(t+2)$, then

$$\phi(\mu_t) - \phi(q^*) \le \frac{4D_\omega(q^*, q_0)}{(t+1)^2}.$$

Proof. For $q^* \in \arg\min_{q\in C} \phi(q)$, we have $\phi(\mu_j) - \phi(q^*) \ge 0$. It then follows from Lemma 2.7 and eq. (2.10) and $\lambda_0 = 1$ that

$$\frac{1}{\lambda_{t-1}^2}\left(\phi(\mu_t) - \phi(q^*)\right) \le D_\omega(q^*, q_0) - D_\omega(q^*, q_t).$$

QED.

2.3.2 A momentum-style algorithm

Our momentum-style algorithm is formally described in Algorithm 2.1. It is equivalent to eq. (2.3) since $Z^\top q_t = Z^\top \nabla\psi(-Zw_t) = -\nabla\widehat{\mathcal{R}}(w_t)/\widehat{\mathcal{R}}(w_t)$.

Here is our main convergence result.

Theorem 2.5. Let w_t and g_t be given by Algorithm 2.1 with $\theta_t = 1$ and $\beta_t = t/(t+1)$, then for all $t \geq 1$,

$$\frac{\min_{1\leq i\leq n} y_i\langle w_t, x_i\rangle}{\|w_t\|_2} \geq \gamma^* - \frac{4\left(1+\ln(n)\right)\left(1+2\ln(t+1)\right)}{\gamma^*(t+1)^2}.$$

Next we prove Theorem 2.5. First, we apply Nesterov's method with the ℓ_1 geometry and entropy regularizer [34, 35] to optimize the dual objective ϕ. Specifically, we run eq. (2.4) with $\|\cdot\| = \|\cdot\|_1$, and $\phi(q) = \left\|Z^\top q\right\|_2^2/2$, and $\omega = \psi^*$ (the entropy function). Concretely, let $\mu_0 = q_0 := (\frac{1}{n}, \ldots, \frac{1}{n})$; for $t \geq 0$, let $\lambda_t, \theta_t \in (0, 1]$, and

$$\begin{aligned}
\nu_t &:= (1-\lambda_t)\mu_t + \lambda_t q_t, \\
q_{t+1} &\propto q_t \odot \exp\left(-\frac{\theta_t}{\lambda_t} ZZ^\top \nu_t\right), \quad q_{t+1} \in \Delta_n, \\
\mu_{t+1} &:= (1-\lambda_t)\mu_t + \lambda_t q_{t+1}.
\end{aligned}$$

By invoking Theorem 2.4, we can show an $O(1/t^2)$ convergence rate on ϕ. However, this rate is not needed in the margin analysis; instead, we should construct the corresponding primal iterates based on the above dual iterates.

Here we construct primal variables $(w_t)_{t=0}^\infty$ such that $\nabla\psi(-Zw_t) = q_t$. (We do not try to make $\nabla\psi(-Zw_t) = \nu_t$ or μ_t, since only q_t is constructed using a mirror descent or dual averaging update.) Let $w_0 := 0$, and for $t \geq 0$, let

$$w_{t+1} := w_t + \frac{\theta_t}{\lambda_t} Z^\top \nu_t. \tag{2.12}$$

We can verify that q_t is indeed the dual variable to w_t, in the sense that $\nabla\psi(-Zw_t) = q_t$: this is true by definition at $t = 0$, since $\nabla\psi(-Zw_0) = \nabla\psi(0) = q_0$. For $t \geq 0$, we have

$$\begin{aligned}
q_{t+1} &\propto q_t \odot \exp\left(-\frac{\theta_t}{\lambda_t} ZZ^\top \nu_t\right) \\
&\propto \exp(-Zw_t) \odot \exp\left(-\frac{\theta_t}{\lambda_t} ZZ^\top \nu_t\right) \\
&= \exp\left(-Z\left(w_t + \frac{\theta_t}{\lambda_t} Z^\top \nu_t\right)\right) = \exp(-Zw_{t+1}).
\end{aligned}$$

Next, we verify that eq. (2.12) is consistent with eq. (2.3) and Algorithm 2.1.

Lemma 2.9. Let w_t be constructed by eq. (2.12). For all $\lambda_t, \theta_t \in (0, 1]$, if $\lambda_0 = 1$, then for all $t \geq 0$,

$$w_{t+1} = w_t - \theta_t \left(g_t - Z^\top q_t \right), \tag{2.13}$$

where $g_0 := 0$, and for $t \geq 1$,

$$g_t := \frac{\lambda_{t-1}(1 - \lambda_t)}{\lambda_t} \left(g_{t-1} - Z^\top q_t \right). \tag{2.14}$$

Specifically, for $\lambda_t = 2/(t+2)$, it holds that

$$\frac{\lambda_{t-1}(1 - \lambda_t)}{\lambda_t} = \frac{t}{t+1}, \quad \text{and} \quad g_t = -\sum_{j=1}^{t} \frac{j}{t+1} Z^\top q_j.$$

Consequently, with $\lambda_t = 2/(t+2)$, the primal iterate defined by eq. (2.12) coincides with the iterate given by Algorithm 2.1 with $\beta_t = t/(t+1)$.

Proof of Lemma 2.9. To prove eq. (2.13), we only need to show that g_t defined by eq. (2.14) satisfies

$$g_t = \frac{w_t - w_{t+1}}{\theta_t} + Z^\top q_t = -Z^\top \left(\frac{1}{\lambda_t} \nu_t - q_t \right).$$

It holds at $t = 0$ by definition, since $\lambda_0 = 1$ and $\nu_0 = q_0$. Moreover,

$$\begin{aligned} \frac{1}{\lambda_{t+1}} \nu_{t+1} - q_{t+1} = \frac{1 - \lambda_{t+1}}{\lambda_{t+1}} \mu_{t+1} &= \frac{1 - \lambda_{t+1}}{\lambda_{t+1}} \left(\nu_t + \lambda_t (q_{t+1} - q_t) \right) \\ &= \frac{\lambda_t (1 - \lambda_{t+1})}{\lambda_{t+1}} \left(\frac{1}{\lambda_t} \nu_t - q_t \right) + \frac{\lambda_t (1 - \lambda_{t+1})}{\lambda_{t+1}} q_{t+1}, \end{aligned}$$

which coincides with the recursive definition of g_t.

For $\lambda_t = 2/(t+2)$, it can be verified directly that $\lambda_t(1 - \lambda_{t+1})/\lambda_{t+1} = (t+1)/(t+2)$. The explicit expression of g_t clearly holds when $t = 0$; for $t \geq 0$,

$$g_{t+1} = \frac{t+1}{t+2} \left(g_t - Z^\top q_{t+1} \right) = -\frac{t+1}{t+2} \sum_{j=1}^{t} \frac{j}{t+1} Z^\top q_j - \frac{t+1}{t+2} Z^\top q_{t+1} = -\sum_{j=1}^{t+1} \frac{j}{t+2} Z^\top q_j.$$

QED.

Next we prove the margin maximization rate. Similarly to the proof in the previous section, we only need to show a lower bound on $-\psi(-Zw_t)$, since $\min_{1 \leq i \leq n} y_i \langle w_t, x_i \rangle \geq -\psi(-Zw_t)$.

Below is our lower bound on $-\psi$; its proof is based on a much finer analysis of dual Nesterov, and uses both primal and dual smoothness.

Lemma 2.10. Let $\theta_t = 1$ for all $t \geq 0$, and $\lambda_0 = 1$, then for all $t \geq 1$,

$$-\psi(-Zw_t) \geq -\psi(-Zw_0) + \frac{1}{2\lambda_{t-1}^2}\left\|Z^\top \mu_t\right\|_2^2 + \sum_{j=1}^{t-1} \frac{1}{2}\left(\frac{1}{\lambda_{j-1}^2} - \frac{1-\lambda_j}{\lambda_j^2}\right)\left\|Z^\top \mu_j\right\|_2^2 + \sum_{j=0}^{t-1} \frac{1}{2\lambda_j}\left\|Z^\top \nu_j\right\|_2^2.$$

Proof. Note that by eq. (2.9),

$$\begin{aligned}\left\langle \nabla\phi(\nu_t), \nu_t - q^*\right\rangle &\leq \frac{1-\lambda_t}{\lambda_t}\left(\phi(\mu_t) - \phi(\nu_t)\right) + \frac{1}{\lambda_t}\left(\phi(\nu_t) - \phi(\mu_{t+1})\right) \\ &\quad + \lambda_t\left(D_{\psi^*}(q^*, q_t) - D_{\psi^*}(q^*, q_{t+1})\right) \\ &= \phi(\nu_t) + \frac{1-\lambda_t}{\lambda_t}\phi(\mu_t) - \frac{1}{\lambda_t}\phi(\mu_{t+1}) + \lambda_t\left(D_{\psi^*}(q^*, q_t) - D_{\psi^*}(q^*, q_{t+1})\right).\end{aligned}$$

Moreover, $\left\langle \nabla\phi(\nu_t), \nu_t\right\rangle = \left\|Z^\top \nu_t\right\|_2^2 = 2\phi(\nu_t)$, and thus

$$\phi(\nu_t) - \left\langle \nabla\phi(\nu_t), q^*\right\rangle \leq \frac{1-\lambda_t}{\lambda_t}\phi(\mu_t) - \frac{1}{\lambda_t}\phi(\mu_{t+1}) + \lambda_t\left(D_{\psi^*}(q^*, q_t) - D_{\psi^*}(q^*, q_{t+1})\right). \quad (2.15)$$

Additionally, let $p_t = -Zw_t$, we have

$$\begin{aligned}D_{\psi^*}(q^*, q_t) - D_{\psi^*}(q^*, q_{t+1}) &= \psi^*(q^*) - \psi^*(q_t) - \langle p_t, q^* - q_t\rangle \\ &\quad - \psi^*(q^*) + \psi^*(q_{t+1}) + \langle p_{t+1}, q^* - q_{t+1}\rangle \\ &= \langle p_t, q_t\rangle - \psi^*(q_t) - \langle p_{t+1}, q_{t+1}\rangle + \psi^*(q_{t+1}) - \langle p_t - p_{t+1}, q^*\rangle \\ &= \psi(p_t) - \psi(p_{t+1}) - \langle p_t - p_{t+1}, q^*\rangle \\ &= \psi(-Zw_t) - \psi(-Zw_{t+1}) - \frac{1}{\lambda_t}\langle \nabla\phi(\nu_t), q^*\rangle \quad (2.16)\end{aligned}$$

Therefore eqs. (2.15) and (2.16) imply

$$\psi(-Zw_t) - \psi(-Zw_{t+1}) \geq \frac{1}{\lambda_t^2}\phi(\mu_{t+1}) - \frac{1-\lambda_t}{\lambda_t^2}\phi(\mu_t) + \frac{1}{\lambda_t}\phi(\nu_t). \quad (2.17)$$

Take the sum of eq. (2.17) from 0 to $t-1$ finishes the proof. QED.

We also need the following bounds on $\|w_t\|_2$.

Lemma 2.11. Let $\theta_t = 1$ for all $t \geq 0$, then

$$\sum_{j=0}^{t-1} \frac{\gamma^*}{\lambda_j} \leq \|w_t\|_2 \leq \sum_{j=0}^{t-1} \frac{1}{\lambda_j} \left\| Z^\top \nu_j \right\|_2 .$$

Proof. The upper bound follows immediately from the triangle inequality. For the lower bound, recall u^* denotes the maximum-margin classifier,

$$\begin{aligned} \|w_t\|_2 \geq \langle w_t, u^* \rangle &= \sum_{j=0}^{t-1} \frac{1}{\lambda_j} \left\langle Z^\top \nu_j, u^* \right\rangle \\ &= \sum_{j=0}^{t-1} \frac{1}{\lambda_j} \left\langle \nu_j, Z u^* \right\rangle \\ &\geq \sum_{j=0}^{t-1} \frac{\gamma^*}{\lambda_j}, \end{aligned}$$

since $\nu_j > 0$, and $\left\|\nu_j\right\|_1 = 1$, and $\langle z_i, u^* \rangle \geq \gamma^*$ for all i. QED.

With Lemmas 2.10 and 2.11, we can prove Theorem 2.5. To illustrate the proof idea, we first show a weaker result which gives $1/t^2$ convergence to $\gamma^*/2$; its proof is also part of the full proof of Theorem 2.5, but much simpler.

Proposition 2.6 (weaker version of Theorem 2.5)**.** With $\theta_t = 1$ and $\lambda_t = 2/(t+2)$, we have

$$\frac{\min_{1 \leq i \leq n} y_i \langle w_t, x_i \rangle}{\|w_t\|_2} \geq \frac{\gamma^*}{2} - \frac{4 \ln(n)}{\gamma^* (t+1)^2}.$$

Proof. With $\lambda_t = 2/(t+2)$, it holds that $\frac{1}{\lambda_{j-1}^2} - \frac{1-\lambda_j}{\lambda_j^2} \geq 0$, therefore we can ignore the $\|Z^\top \mu_j\|_2$ terms in Lemma 2.10 and get

$$-\psi(-Z w_t) \geq -\psi(-Z w_0) + \sum_{j=0}^{t-1} \frac{1}{2\lambda_j} \left\| Z^\top \nu_j \right\|_2^2. \tag{2.18}$$

Then eq. (2.18) and Lemma 2.11 imply

$$\frac{\psi(-Z w_0) - \psi(-Z w_t)}{\|w_t\|_2} \geq \frac{\sum_{j=0}^{t-1} \frac{1}{2\lambda_j} \left\| Z^\top \nu_j \right\|_2^2}{\sum_{j=0}^{t-1} \frac{1}{\lambda_j} \left\| Z^\top \nu_j \right\|_2} \geq \frac{\gamma^*}{2}, \tag{2.19}$$

since $\left\| Z^\top \nu_j \right\|_2 \geq \gamma^*$ (cf. Lemma 2.3). On the other hand, Lemma 2.11 and $\lambda_t = 2/(t+2)$

imply

$$\|w_t\|_2 \geq \sum_{j=0}^{t-1} \frac{\gamma^*}{\lambda_j} \geq \frac{\gamma^*(t+1)^2}{4},$$

and thus

$$\frac{\psi(-Zw_0)}{\|w_t\|_2} = \frac{\ln(n)}{\|w_t\|_2} \leq \frac{4\ln(n)}{\gamma^*(t+1)^2}. \tag{2.20}$$

It then follows from eqs. (2.19) and (2.20) that

$$\frac{\min_{1\leq i\leq n} y_i\langle w_t, x_i\rangle}{\|w_t\|_2} \geq \frac{-\psi(-Zw_t)}{\|w_t\|_2} \geq \frac{\gamma^*}{2} - \frac{4\ln(n)}{\gamma^*(t+1)^2}.$$

QED.

To prove the full version of Theorem 2.5, we need the following result which gives an alternative characterization of w_t using μ_j.

Lemma 2.12. Let $\theta_t = 1$, we have

$$w_t = Z^\top q_0 - Z^\top q_t + \sum_{j=0}^{t-1} \frac{1}{\lambda_j} Z^\top \mu_{j+1},$$

and if $\lambda_t = 2/(t+2)$, then

$$\frac{1}{2\lambda_{t-1}^2}\left\|Z^\top \mu_t\right\|_2^2 + \sum_{j=1}^{t-1} \frac{1}{2}\left(\frac{1}{\lambda_{j-1}^2} - \frac{1-\lambda_j}{\lambda_j^2}\right)\left\|Z^\top \mu_j\right\|_2^2 \geq \sum_{j=0}^{t-1} \frac{1}{2\lambda_j}\left\|Z^\top \mu_{j+1}\right\|_2^2 - 2\ln(n)\ln(t+1).$$

Proof. Note that by construction, $\frac{1}{\lambda_t}\nu_t = \frac{1-\lambda_t}{\lambda_t}\mu_t + q_t$, and thus

$$\begin{aligned} w_t = \sum_{j=0}^{t-1} Z^\top\left(\frac{1}{\lambda_j}\nu_j\right) &= \sum_{j=0}^{t-1} Z^\top\left(\frac{1-\lambda_j}{\lambda_j}\mu_j + q_j\right) \\ &= Z^\top q_0 - Z^\top q_t + \sum_{j=0}^{t-1} Z^\top\left(\frac{1-\lambda_j}{\lambda_j}\mu_j + q_{j+1}\right) \\ &= Z^\top q_0 - Z^\top q_t + \sum_{j=0}^{t-1} Z^\top\left(\frac{1}{\lambda_j}\mu_{j+1}\right). \end{aligned}$$

On the second claim, note that

$$\begin{aligned}
&\frac{1}{2\lambda_{t-1}^2}\left\|Z^\top \mu_t\right\|_2^2 + \sum_{j=1}^{t-1}\frac{1}{2}\left(\frac{1}{\lambda_{j-1}^2} - \frac{1-\lambda_j}{\lambda_j^2}\right)\left\|Z^\top \mu_j\right\|_2^2 \\
\geq\ & \frac{1}{2\lambda_{t-1}^2}(\gamma^*)^2 + \sum_{j=1}^{t-1}\frac{1}{2}\left(\frac{1}{\lambda_{j-1}^2} - \frac{1-\lambda_j}{\lambda_j^2}\right)(\gamma^*)^2 \\
=\ & \sum_{j=0}^{t-1}\frac{1}{2\lambda_j}(\gamma^*)^2.
\end{aligned}$$

Moreover, Theorem 2.4 implies

$$\frac{1}{2\lambda_j}\left(\left\|Z^\top \mu_{j+1}\right\|_2^2 - (\gamma^*)^2\right) \leq \lambda_j D_{\psi^*}(q^*, q_0) \leq \lambda_j \ln(n),$$

and thus

$$\begin{aligned}
&\frac{1}{2\lambda_{t-1}^2}\left\|Z^\top \mu_t\right\|_2^2 + \sum_{j=1}^{t-1}\frac{1}{2}\left(\frac{1}{\lambda_{j-1}^2} - \frac{1-\lambda_j}{\lambda_j^2}\right)\left\|Z^\top \mu_j\right\|_2^2 \\
\geq\ & \sum_{j=0}^{t-1}\frac{1}{2\lambda_j}(\gamma^*)^2 \\
=\ & \sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top \mu_{j+1}\right\|_2^2 - \sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left(\left\|Z^\top \mu_{j+1}\right\|_2^2 - (\gamma^*)^2\right) \\
\geq\ & \sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top \mu_{j+1}\right\|_2^2 - \ln(n)\sum_{j=0}^{t-1}\lambda_j.
\end{aligned}$$

Finally, note that

$$\sum_{j=0}^{t-1}\lambda_j = \sum_{j=0}^{t-1}\frac{2}{j+2} \leq 2\ln(t+1),$$

the proof is done. QED.

Now we can prove the full version of Theorem 2.5.

Proof of Theorem 2.5. Lemmas 2.10 and 2.12 imply

$$\psi(-Zw_0) - \psi(-Zw_t) \geq \sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top \mu_{j+1}\right\|_2^2 + \sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top \nu_j\right\|_2^2 - 2\ln(n)\ln(t+1).$$

Therefore

$$\begin{aligned}
&\frac{\min_{1\le i\le n} y_i\langle w_t, x_i\rangle}{\|w_t\|_2} \\
\ge\ & \frac{\psi(-Zw_0)-\psi(-Zw_t)}{\|w_t\|_2} - \frac{\psi(-Zw_0)}{\|w_t\|_2} \\
\ge\ & \frac{\sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2^2}{\|w_t\|_2} + \frac{\sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top\nu_j\right\|_2^2}{\|w_t\|_2} - \frac{2\ln(n)\ln(t+1)}{\|w_t\|_2} - \frac{\ln(n)}{\|w_t\|_2} \\
=\ & \frac{\sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2^2}{\|w_t\|_2} + \frac{\sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top\nu_j\right\|_2^2}{\|w_t\|_2} - \frac{\ln(n)\left(1+2\ln(t+1)\right)}{\|w_t\|_2}. \qquad (2.21)
\end{aligned}$$

By the triangle inequality and the alternative characterization of w_t in Lemma 2.12, we have

$$\|w_t\|_2 \le \left\|Z^\top q_0\right\|_2 + \left\|Z^\top q_t\right\|_2 + \sum_{j=0}^{t-1}\frac{1}{\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2 \le 2 + \sum_{j=0}^{t-1}\frac{1}{\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2.$$

Therefore

$$\begin{aligned}
\frac{\sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2^2}{\|w_t\|_2} &\ge \frac{\gamma^*\sum_{j=0}^{t-1}\frac{1}{2\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2}{2+\sum_{j=0}^{t-1}\frac{1}{\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2} \\
&= \frac{\gamma^*}{2}\left(1-\frac{2}{2+\sum_{j=0}^{t-1}\frac{1}{\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2}\right) \\
&\ge \frac{\gamma^*}{2}\left(1-\frac{2}{\sum_{j=0}^{t-1}\frac{1}{\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2}\right) \ge \frac{\gamma^*}{2}\left(1-\frac{8}{\gamma^*(t+1)^2}\right),
\end{aligned}$$

where we use $\sum_{j=0}^{t-1}\frac{1}{\lambda_j}\left\|Z^\top\mu_{j+1}\right\|_2 \ge \sum_{j=0}^{t-1}\frac{\gamma^*}{\lambda_j} \ge \frac{\gamma^*(t+1)^2}{4}$. The remaining part of eq. (2.21) can be handled in the same way as in the proof of Proposition 2.6. QED.

2.4 GENERAL DECREASING LOSSES

In previous sections, we focus on exponentially-tailed losses, such as the exponential loss or logistic loss. Here we further consider general convex decreasing losses. We show that it is still possible to characterize the implicit bias of GD, in terms of the *regularization path*: given $B \ge 0$, let $\bar{w}(B)$ denote the regularized solution with ℓ_2 norm bounded by B, concretely

$$\bar{w}(B) := \arg\min_{\|w\|_2\le B} \widehat{\mathcal{R}}(w), \qquad (2.22)$$

and the regularization path denotes the curve followed by $\bar{w}$ as B varies, meaning $(\bar{w}(B))_{B\geq 0}$. (Choosing regularized rather than constrained solutions does not change our results regarding the regularization path; moreover, in either case, the paths are algorithm-independent.) Then under some mild conditions, we can show that $\lim_{t\to\infty} \frac{w_t}{\|w_t\|_2} = \lim_{B\to\infty} \frac{\bar{w}(B)}{B}$ whenever either limit exists.

We also show that different losses can induce very different implicit biases: exponentially-tailed losses all converge to the maximum-margin direction, but polynomially-tailed losses (cf. eq. (2.41)) may converge to a direction with a poor margin.

The contents in this section are based on [22].

2.4.1 Convergence of GD implies convergence of regularization path

In this subsection we show one direction of the equivalence, which holds in a more general setting. Given a differentiable convex function $f : \mathbb{R}^d \to \mathbb{R}$ (not necessarily the empirical risk) and an ℓ_2-norm bound B, the regularized solution is defined as

$$\bar{w}(B) := \arg\min_{\|w\|_2\leq B} f(w).$$

Note that $\bar{w}(B)$ is not unique in general, but we still have $\lim_{B\to\infty} f\left(\bar{w}(B)\right) = \inf_{w\in\mathbb{R}^d} f(w)$, as is often the case when working with unregularized losses. In this paper we are particularly interested in the case where the infimum of f is not attained. In that case $\bar{w}(B)$ is uniquely defined, because the set of minimizers is convex and contained in the surface of the ℓ_2 ball, and thus consists of exactly one point due to the curvature of ℓ_2 balls. A particular case is an empirical risk with a nonempty separable part, which has been studies in previous sections.

We consider GD with a constant learning rate η. Its basic properties are summarized in Lemma 2.13. If there exists a small step size which ensures decreasing function values, then gradient descent on f can minimize the function value to its infimum; moreover, if the infimum of f is not attained, then gradient descent iterates go to infinity.

Lemma 2.13. Given a convex differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, suppose the step size η satisfies

$$f(w_{t+1}) - f(w_t) \leq -\frac{\eta}{2}\left\|\nabla f(w_t)\right\|_2^2 \tag{2.23}$$

for all $t \geq 0$. Then for any $w \in \mathbb{R}^d$,

$$\|w_{t+1} - w\|_2^2 \leq \|w_t - w\|_2^2 + 2\eta\left(f(w) - f(w_{t+1})\right), \tag{2.24}$$

and thus $\|w_{t+1} - w\|_2 \leq \|w_t - w\|_2$ as long as $f(w) \leq f(w_{t+1})$. Consequently,

$$\lim_{t\to\infty} f(w_t) = \inf_{w\in\mathbb{R}^d} f(w),$$

which implies $\lim_{t\to\infty} \|w_t\|_2 = \infty$ if the infimum of f is not attained.

Proof. For any $w \in \mathbb{R}^d$, it holds that

$$\begin{aligned}
\|w_{t+1} - w\|_2^2 &= \|w_t - w\|_2^2 - 2\eta \left\langle \nabla f(w_t), w_t - w \right\rangle + \eta^2 \left\|\nabla f(w_t)\right\|_2^2 \\
&= \|w_t - w\|_2^2 + 2\eta \left\langle \nabla f(w_t), w - w_t \right\rangle + 2\eta \cdot \frac{\eta}{2} \left\|\nabla f(w_t)\right\|_2^2 \\
&\leq \|w_t - w\|_2^2 + 2\eta \left(f(w) - f(w_t)\right) + 2\eta \left(f(w_t) - f(w_{t+1})\right) \\
&= \|w_t - w\|_2^2 + 2\eta \left(f(w) - f(w_{t+1})\right). \qquad (2.25)
\end{aligned}$$

On the third line we use the convexity of f and eq. (2.23).

Since $f(w_t)$ is nondecreasing, $\lim_{t\to\infty} f(w_t)$ exists. Suppose $\lim_{t\to\infty} f(w_t) > \inf_{w\in\mathbb{R}^d} f(w)$. Let $\bar{w} \in \mathbb{R}^d$ satisfy $f(\bar{w}) < \lim_{t\to\infty} f(w_t) - \epsilon$ for some $\epsilon > 0$. It follows from eq. (2.25) that $\|w_{t+1} - \bar{w}\|_2^2 \leq \|w_t - \bar{w}\|_2^2 - 2\eta\epsilon$ for any t, which implies $\|w_{t+1} - \bar{w}\|_2^2 \to -\infty$, which is a contradiction. QED.

Remark 2.1. The step size condition in eq. (2.23) holds if f is (globally) β-smooth and $\eta \leq 1/\beta$. There are also standard situations where f merely obeys local smoothness over its sublevel sets; see for example Lemma 3.16, which considers empirical risk minimization with the exponential loss.

Below is our main result of this section.

Theorem 2.7. Suppose f is convex, differentiable, bounded below by 0, and its infimum is not attained, and the step size η satisfies eq. (2.23) and $\eta \leq 1/\left(2f(w_0)\right)$. If $\lim_{t\to\infty} w_t/\|w_t\|_2 = \bar{u}$ for some unit vector $\bar{u}$, then also $\lim_{B\to\infty} \bar{w}(B)/B = \bar{u}$.

Proof. The first observation is that for any $\epsilon > 0$, there exists $B_1(\epsilon) > 0$, such that for any GD iterate w_t with $\|w_t\|_2 > B_1(\epsilon)$, it holds that $\left\| \frac{w_t}{\|w_t\|_2} - \bar{u} \right\|_2 < \epsilon$. Given any ϵ, by our assumption, there exists t_1 such that $\left\| \frac{w_t}{\|w_t\|_2} - \bar{u} \right\|_2 < \epsilon$ for any $t > t_1$. It is enough to let $B_1(\epsilon) = \max_{0\leq t\leq t_1} \|w_t\|_2 + 1$.

Then we show that $\lim_{B\to\infty} \left\langle \bar{w}(B), \bar{u} \right\rangle \to \infty$. If this is not true, then there exists a constant $C > 0$ such that there exists arbitrarily large B with $\left\langle \bar{w}(B), \bar{u} \right\rangle < C$. Choose B_2

such that

$$B_2 > \max\left\{5\left(\|w_0\|_2 + C + 1\right), B_1\left(\frac{1}{4}\right) + 1\right\}, \quad \text{and} \quad \left\langle \bar{w}(B_2), \bar{u}\right\rangle < C.$$

Let t_2 denote the first step such that $\|w_{t_2}\|_2 > B_2 - 1$. Since $B_2 - 1 > \|w_0\|_2$, we have $t_2 > 0$. Moreover, the conditions of Theorem 2.7 (i.e., eq. (2.23) and $\eta \leq 1/\left(2f(w_0)\right)$) implies

$$\begin{aligned}\|w_{t_2} - w_{t_2-1}\|_2 = \eta\left\|\nabla f(w_{t_2-1})\right\|_2 =& \sqrt{\eta^2\left\|\nabla f(w_{t_2-1})\right\|_2^2} \\ &\leq \sqrt{2\eta\left(f(w_{t_2-1}) - f(w_{t_2})\right)} \\ &\leq \sqrt{2\eta f(w_0)} \leq 1.\end{aligned} \tag{2.26}$$

Therefore from the definition of t_2,

$$\|w_{t_2}\|_2 \leq \|w_{t_2-1}\|_2 + \|w_{t_2} - w_{t_2-1}\|_2 \leq B_2 - 1 + 1 = B_2.$$

By the definition of t_2 and $\bar{w}(B_2)$, we have $f\left(\bar{w}(B_2)\right) \leq f(w_t)$ for any $t \leq t_2$. Using eq. (2.24), we can show that

$$\left\|w_{t_2} - \bar{w}(B_2)\right\|_2 \leq \left\|w_0 - \bar{w}(B_2)\right\|_2. \tag{2.27}$$

On one hand,

$$\left\|w_0 - \bar{w}(B_2)\right\|_2 \leq \|w_0\|_2 + \left\|\bar{w}(B_2)\right\|_2 = \|w_0\|_2 + B_2. \tag{2.28}$$

On the other hand,

$$\begin{aligned}\left\|w_{t_2} - \bar{w}(B_2)\right\|_2^2 &= \|w_{t_2}\|_2^2 + B_2^2 - 2\left\langle w_{t_2}, \bar{w}(B_2)\right\rangle \\ &= \|w_{t_2}\|_2^2 + B_2^2 - 2\|w_{t_2}\|_2\left\langle \frac{w_{t_2}}{\|w_{t_2}\|_2}, \bar{w}(B_2)\right\rangle \\ &> (B_2 - 1)^2 + B_2^2 - 2\|w_{t_2}\|_2\left\langle \frac{w_{t_2}}{\|w_{t_2}\|_2}, \bar{w}(B_2)\right\rangle.\end{aligned}$$

By the definition of t_2 and B_2, we have

$$\|w_{t_2}\|_2 > B_2 - 1 > B_1\left(\frac{1}{4}\right),$$

and thus $\left\| \frac{w_{t_2}}{\|w_{t_2}\|_2} - \bar{u} \right\|_2 < 1/4$. As a result,

$$\left\langle \frac{w_{t_2}}{\|w_{t_2}\|_2}, \bar{w}(B_2) \right\rangle < \left\langle \bar{u}, \bar{w}(B_2) \right\rangle + \frac{1}{4} B_2 < C + \frac{1}{4} B_2,$$

and

$$\begin{aligned} \left\| w_{t_2} - \bar{w}(B_2) \right\|_2^2 &> (B_2 - 1)^2 + B_2^2 - 2\|w_{t_2}\|_2 C - \frac{1}{2}\|w_{t_2}\|_2 B_2 \\ &\geq (B_2 - 1)^2 + B_2^2 - 2CB_2 - \frac{1}{2} B_2^2 > \frac{3}{2} B_2^2 - 2CB_2 - 2B_2. \end{aligned} \tag{2.29}$$

Combining eqs. (2.27) to (2.29) gives

$$\frac{3}{2} B_2^2 - 2CB_2 - 2B_2 < \|w_0\|_2^2 + 2\|w_0\|_2 B_2 + B_2^2,$$

which implies

$$B_2 < 4\left(\|w_0\|_2 + C + 1\right) + \frac{2\|w_0\|_2^2}{B_2} < 4\left(\|w_0\|_2 + C + 1\right) + \|w_0\|_2 < 5\left(\|w_0\|_2 + C + 1\right),$$

a contradiction.

Next we prove the claim that $\lim_{B \to \infty} \bar{w}(B)/B = \bar{u}$. If this is not true, then there exists $\delta > 0$, such that there exists arbitrarily large B with $\left\| \frac{\bar{w}(B)}{B} - \bar{u} \right\|_2 > \delta$. Choose B_4 such that

$$\left\| \frac{\bar{w}(B_4)}{B_4} - \bar{u} \right\|_2 > \delta, \quad \text{and} \quad \left\langle \bar{w}(B_4), \bar{u} \right\rangle > B_1 \left(\frac{\delta^3}{32} \right) + \|w_0\|_2 + 1, \quad \text{and} \quad B_4 > \frac{32}{\delta^3}.$$

Let $B_3 := \left\langle \bar{w}(B_4), \bar{u} \right\rangle$. By geometric arguments, we have

$$\left\| \bar{w}(B_4) - B_4 \bar{u} \right\|_2 - \left\| \bar{w}(B_4) - B_3 \bar{u} \right\|_2 > \frac{B_4 \delta^3}{8}. \tag{2.30}$$

Let t_3 denote the first step such that $\|w_{t_3}\|_2 > B_3 - 1$. Since $B_3 - 1 > \|w_0\|_2$, we have $t_3 > 0$, and similar to eq. (2.26) we can show that $\|w_{t_3}\|_2 \leq B_3$. Since $B_3 - 1 > B_1(\delta^3/32)$, we have $\left\| \frac{w_{t_3}}{\|w_{t_3}\|_2} - \bar{u} \right\|_2 < \delta^3/32$. As a result,

$$\begin{aligned} \|w_{t_3} - B_3 \bar{u}\|_2 &\leq \left\| w_{t_3} - \|w_{t_3}\|_2 \bar{u} \right\|_2 + \left\| \|w_{t_3}\|_2 \bar{u} - B_3 \bar{u} \right\|_2 \\ &\leq \|w_{t_3}\|_2 \frac{\delta^3}{32} + 1 \leq \frac{B_3 \delta^3}{32} + 1 \leq \frac{B_4 \delta^3}{32} + 1. \end{aligned} \tag{2.31}$$

Similarly, let t_4 denote the first step such that $\|w_{t_4}\|_2 > B_4 - 1$, we can show that $\|w_{t_4}\|_2 \le B_4$, and

$$\|w_{t_4} - B_4\bar{u}\|_2 \le \frac{B_4\delta^3}{32} + 1. \tag{2.32}$$

Combining eqs. (2.30) to (2.32) gives

$$\begin{aligned}
&\left\|\bar{w}(B_4) - w_{t_4}\right\|_2 - \left\|\bar{w}(B_4) - w_{t_3}\right\|_2 \\
\ge &\left\|\bar{w}(B_4) - B_4\bar{u}\right\|_2 - \|B_4\bar{u} - w_{t_4}\|_2 - \left\|\bar{w}(B_4) - B_3\bar{u}\right\|_2 - \|B_3\bar{u} - w_{t_3}\|_2 \\
\ge &\frac{B_4\delta^3}{8} - \frac{B_4\delta^3}{32} - 1 - \frac{B_4\delta^3}{32} - 1 \\
= &\frac{B_4\delta^3}{16} - 2 > 0.
\end{aligned} \tag{2.33}$$

On the other hand, using eq. (2.30) and the triangle inequality,

$$B_4 - B_3 = \|B_4\bar{u} - B_3\bar{u}\|_2 \ge \left\|\bar{w}(B_4) - B_4\bar{u}\right\|_2 - \left\|\bar{w}(B_4) - B_3\bar{u}\right\|_2 > \frac{B_4\delta^3}{8} > 4,$$

and thus $t_4 > t_3$. Since $\|w_{t_4}\|_2 \le B_4$, by the definition of t_4 and $\bar{w}(B_4)$, we have $f\left(\bar{w}(B_4)\right) \le f(w_t)$ for any $t \le t_4$. Since $t_3 < t_4$, eq. (2.24) implies $\left\|\bar{w}(B_4) - w_{t_4}\right\|_2 \le \left\|\bar{w}(B_4) - w_{t_3}\right\|_2$, which contradicts eq. (2.33). QED.

2.4.2 Convergence of regularization path implies Convergence of GD

From now on, we consider empirical risk minimization with binary classification. The training data is still assumed to be linearly separable, and the loss function ℓ is assumed to be convex, differentiable, and strictly decreasing to 0, such as the exponential loss or logistic loss.

Linear separability and a strictly decreasing loss imply that the infimum of $\widehat{\mathcal{R}}$ is not attained, and thus Theorem 2.7 can be applied. However, we can show a stronger result: the GD path converges to a direction if and only if the regularization path converges to (the same) direction.

Theorem 2.8. Suppose the step size satisfies $\eta \le 1/\left(2\widehat{\mathcal{R}}(w_0)\right)$ and

$$\widehat{\mathcal{R}}(w_{t+1}) - \widehat{\mathcal{R}}(w_t) \le -\frac{\eta}{2}\left\|\nabla\widehat{\mathcal{R}}(w_t)\right\|_2^2 \tag{2.34}$$

for all $t \ge 0$. Then $\lim_{t\to\infty} w_t/\|w_t\|_2$ exists if and only if $\lim_{B\to\infty} \bar{w}(B)/B$ exists, and when

they exist they are the same.

The "if" part of Theorem 2.8 follows directly from Theorem 2.7. Next we give a proof of the "only if" part of Theorem 2.8.

In the remainder of this section, we assume that $\lim_{B\to\infty} \bar{w}(B)/B = \bar{u}$ for some unit vector $\bar{u}$, and define its margin as

$$\bar{\gamma} := \min_{1\le i\le n} y_i \langle \bar{u}, x_i \rangle.$$

Also recall that γ^* denotes the maximum margin and u^* denotes the maximum-margin solution. We first show that $\bar{\gamma}$ is always positive.

Lemma 2.14. It holds that $\bar{\gamma} \ge (\gamma^*)^2/(2n) > 0$.

Remark 2.2. Lemma 2.14 gives a worst-case lower bound on $\bar{\gamma}$ which holds for an arbitrary decreasing loss. The proof technique can also be adapted to a specific loss function. For example, if the loss function has a polynomial tail az^{-b}, then $\lim_{B\to\infty} \bar{w}(B)/B$ exists (cf. Proposition 2.9), and we can prove an $\Omega(n^{-1/(b+1)})$ lower bound on $\bar{\gamma}$. Moreover, there exists a dataset on which this lower bound is tight (cf. Proposition 2.10).

To prove Lemma 2.14, we first need the following result.

Lemma 2.15. It holds that

$$\frac{\bar{w}(B)}{B} = -\frac{\nabla \widehat{\mathcal{R}}\left(\bar{w}(B)\right)}{\left\|\nabla \widehat{\mathcal{R}}\left(\bar{w}(B)\right)\right\|_2}.$$

Conversely, if $\|w\|_2 = B$ and $w/B = -\nabla\widehat{\mathcal{R}}(w)/\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2$, then $w = \bar{w}(B)$.

Proof. By the first order optimality conditions, $w = \bar{w}(B)$ if and only if for any w' with $\|w'\|_2 \le B$, it holds that

$$\left\langle \nabla\widehat{\mathcal{R}}(w), w' - w \right\rangle \ge 0. \tag{2.35}$$

Since the infimum of $\widehat{\mathcal{R}}$ is not attained, the gradient $\nabla\widehat{\mathcal{R}}(w)$ is always nonzero. The structure of the ℓ_2 ball implies that eq. (2.35) holds if and only if $\|w\|_2 = B$ and $w/B = -\nabla\widehat{\mathcal{R}}(w)/\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2$. QED.

Now we can prove Lemma 2.14.

Proof of Lemma 2.14. Since $\bar{w}(B)/B \to \bar{u}$, the margin of $\bar{w}(B)/B$ converges to the margin of $\bar{u}$. For large enough B, the risk $\widehat{\mathcal{R}}\left(\bar{w}(B)\right) \leq \ell(0)/n$, which implies $\bar{w}(B)/B$ has a nonnegative margin, and thus $\bar{u}$ also has a nonnegative margin.

The proof of Lemma 2.14 is by contradiction. Given $\epsilon := (\gamma^*)^2/(2n)$, suppose there exists $B_0 > 0$, such that for any $B \geq B_0$, the margin of $\bar{w}(B)/B$ is no larger than ϵ. We will derive a contradiction, which implies that the margin of $\bar{u}$ is at least $(\gamma^*)^2/(2n)$.

For any $B > 0$, Lemma 2.15 ensures that

$$-\left\langle \frac{\bar{w}(B)}{B}, \nabla\widehat{\mathcal{R}}\left(\bar{w}(B)\right)\right\rangle = \left\|\nabla\widehat{\mathcal{R}}\left(\bar{w}(B)\right)\right\|_2. \tag{2.36}$$

For simplicity, let $z_i := y_i x_i$. The left hand side of eq. (2.36) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n} -\ell'\left(\left\langle \bar{w}(B), z_i\right\rangle\right)\left\langle \frac{\bar{w}(B)}{B}, z_i\right\rangle, \tag{2.37}$$

while the right hand side of eq. (2.36) can be bounded below as

$$\left\|\nabla\widehat{\mathcal{R}}\left(\bar{w}(B)\right)\right\|_2 \geq \left\langle -\nabla\widehat{\mathcal{R}}\left(\bar{w}(B)\right), u^*\right\rangle \geq \frac{1}{n}\sum_{i=1}^{n} -\ell'\left(\left\langle \bar{w}(B), z_i\right\rangle\right)\gamma^*. \tag{2.38}$$

Let H denote the set of data points on which $\bar{w}(B)/B$ has margin larger than γ^*, and suppose without loss of generality that $\bar{w}(B)/B$ achieves its minimum margin on z_1. It follows from eqs. (2.36) to (2.38) that

$$\begin{aligned}\sum_{z_i\in H} -\ell'\left(\left\langle \bar{w}(B), z_i\right\rangle\right)\left(\left\langle \frac{\bar{w}(B)}{B}, z_i\right\rangle - \gamma^*\right) &\geq \sum_{z_i\notin H} -\ell'\left(\left\langle \bar{w}(B), z_i\right\rangle\right)\left(\gamma^* - \left\langle \frac{\bar{w}(B)}{B}, z_i\right\rangle\right) \\ &\geq -\ell'\left(\left\langle \bar{w}(B), z_1\right\rangle\right)\left(\gamma^* - \left\langle \frac{\bar{w}(B)}{B}, z_1\right\rangle\right).\end{aligned} \tag{2.39}$$

Now consider $B \geq B_0$, which implies $\left\langle \bar{w}(B)/B, z_1\right\rangle \leq \epsilon$. Since $\epsilon < \gamma^*/2$, and $\|z_i\|_2 \leq 1$, eq. (2.39) implies $-n\ell'(B\gamma^*) \geq -\ell'(B\epsilon)(\gamma^* - \epsilon) \geq -\ell'(B\epsilon)\gamma^*/2$, and thus

$$\frac{-\ell'(B\epsilon)}{-\ell'(B\gamma^*)} \leq \frac{2n}{\gamma^*} \tag{2.40}$$

for all $B \geq B_0$. Let $\alpha := B_0\epsilon = B_0(\gamma^*)^2/(2n)$, and $\lambda := 2n/\gamma^*$, then it means for all $z \geq \alpha$, we have $-\ell'(z) \leq -\ell'(\lambda z)\lambda$.

Therefore for any $k \geq 1$, we have

$$\int_{\alpha\lambda^k}^{\alpha\lambda^{k+1}} -\ell'(z)\,\mathrm{d}z = \int_{\alpha\lambda^{k-1}}^{\alpha\lambda^k} -\ell'(\lambda y)\lambda\,\mathrm{d}y \geq \int_{\alpha\lambda^{k-1}}^{\alpha\lambda^k} -\ell'(y)\,\mathrm{d}y,$$

where eq. (2.40) is used. By induction, we have

$$\int_{\alpha\lambda^k}^{\alpha\lambda^{k+1}} -\ell'(z)\,\mathrm{d}z \geq \int_{\alpha}^{\alpha\lambda} -\ell'(z)\,\mathrm{d}z > 0.$$

As a result,

$$\int_{\alpha}^{\infty} -\ell'(z)\,\mathrm{d}z = \infty,$$

which is contradiction, since $\int_\alpha^\infty -\ell'(z)\,\mathrm{d}z = \ell(\alpha)$ should be finite. QED.

Next we can show that to minimize the risk, it is almost optimal to move along the direction of $\bar{u}$, thanks to its positive margin.

Lemma 2.16. Given any $\alpha > 0$, there exists $\rho(\alpha) > 0$, such that for any w with $\|w\|_2 > \rho(\alpha)$, it holds that

$$\widehat{\mathcal{R}}\left((1+\alpha)\|w\|_2\bar{u}\right) \leq \widehat{\mathcal{R}}(w).$$

Proof of Lemma 2.16. Since $\lim_{B\to\infty} \bar{w}(B)/B = \bar{u}$, we can choose $\rho(\alpha)$ large enough such that for any w with $\|w\|_2 > \rho(\alpha)$, it holds that

$$\left\|\bar{w}\left(\|w\|\right)/\|w\|_2 - \bar{u}\right\| \leq \alpha\bar{\gamma}.$$

In this case, for any $1 \leq i \leq n$,

$$\begin{aligned} y_i\left\langle \bar{w}\left(\|w\|_2\right), x_i\right\rangle &= y_i\left\langle \bar{w}\left(\|w\|_2\right) - \|w\|_2\bar{u}, x_i\right\rangle + y_i\left\langle \|w\|_2\bar{u}, x_i\right\rangle \\ &\leq \alpha\bar{\gamma}\|w\|_2 + y_i\left\langle \|w\|_2\bar{u}, x_i\right\rangle \\ &\leq y_i\left\langle (1+\alpha)\|w\|_2\bar{u}, x_i\right\rangle. \end{aligned}$$

As a result,

$$\widehat{\mathcal{R}}\left((1+\alpha)\|w\|_2\bar{u}\right) \leq \widehat{\mathcal{R}}\left(\bar{w}\left(\|w\|_2\right)\right) \leq \widehat{\mathcal{R}}(w).$$

QED.

Now we are ready to prove the "only if" part of Theorem 2.8.

Proof of Theorem 2.8, the "only if" part. Given any $\epsilon \in (0,1)$, let α satisfy $1/(1+\alpha) = 1-\epsilon$ (i.e., let $\alpha = \epsilon/(1-\epsilon)$). Since $\lim_{t\to\infty} \|w_t\|_2 = \infty$, we can choose a step t_0 such that for any $t \geq t_0$, it holds that $\|w_t\|_2 > \max\left\{\rho(\alpha), 1\right\}$, where ρ is given by Lemma 2.16.

Now for any $t \geq t_0$, using the convexity of $\widehat{\mathcal{R}}$ and Lemma 2.16, we have

$$\left\langle \nabla\widehat{\mathcal{R}}(w_t), w_t - (1+\alpha)\|w_t\|_2\bar{u} \right\rangle \geq \widehat{\mathcal{R}}(w_t) - \widehat{\mathcal{R}}\left((1+\alpha)\|w_t\|_2\bar{u}\right) \geq 0,$$

meaning

$$\left\langle \nabla\widehat{\mathcal{R}}(w_t), w_t \right\rangle \geq (1+\alpha)\|w_t\|_2 \left\langle \nabla\widehat{\mathcal{R}}(w_t), \bar{u} \right\rangle.$$

Consequently,

$$\begin{aligned}
\langle w_{t+1} - w_t, \bar{u}\rangle = \left\langle -\eta\nabla\widehat{\mathcal{R}}(w_t), \bar{u} \right\rangle &\geq \left\langle -\eta\nabla\widehat{\mathcal{R}}(w_t), w_t \right\rangle \frac{1}{(1+\alpha)\|w_t\|_2} \\
&= \langle w_{t+1} - w_t, w_t\rangle \frac{1}{(1+\alpha)\|w_t\|_2} \\
&= \left(\frac{1}{2}\|w_{t+1}\|_2^2 - \frac{1}{2}\|w_t\|_2^2 - \frac{1}{2}\|w_{t+1} - w_t\|_2^2\right) \frac{1}{(1+\alpha)\|w_t\|_2}.
\end{aligned}$$

On one hand, we have

$$\left(\frac{1}{2}\|w_{t+1}\|_2^2 - \frac{1}{2}\|w_t\|_2^2\right) / \|w_t\|_2 \geq \|w_{t+1}\|_2 - \|w_t\|_2.$$

On the other hand, using the step size condition in eq. (2.34), we have

$$\frac{\|w_{t+1} - w_t\|_2^2}{2(1+\alpha)\|w_t\|_2} \leq \frac{\|w_{t+1} - w_t\|_2^2}{2} = \frac{\eta^2 \left\|\nabla\widehat{\mathcal{R}}(w_t)\right\|_2^2}{2} \leq \eta\left(\widehat{\mathcal{R}}(w_t) - \widehat{\mathcal{R}}(w_{t+1})\right).$$

As a result,

$$\langle w_t - w_{t_0}, \bar{u}\rangle \geq \frac{\|w_t\|_2 - \|w_{t_0}\|_2}{1+\alpha} - \eta\widehat{\mathcal{R}}(w_{t_0}) = (1-\epsilon)\left(\|w_t\| - \|w_{t_0}\|_2\right) - \eta\widehat{\mathcal{R}}(w_{t_0}),$$

meaning

$$\left\langle \frac{w_t}{\|w_t\|_2}, \bar{u} \right\rangle \geq 1 - \epsilon + \frac{\langle w_{t_0}, \bar{u}\rangle - (1-\epsilon)\|w_{t_0}\|_2 - \eta\widehat{\mathcal{R}}(w_{t_0})}{\|w_t\|_2}.$$

Consequently,

$$\liminf_{t\to\infty} \left\langle \frac{w_t}{\|w_t\|_2}, \bar{u} \right\rangle \geq 1 - \epsilon.$$

Since ϵ is arbitrary, we get $w_t/\|w_t\|_2 \to \bar{u}$. QED.

2.4.3 Applications

Here we briefly review some applications of Theorem 2.8; the detailed proofs can be found in [22].

Theorem 2.8 says that the GD path and regularization path converge to the same direction if either of them converges to a direction. Moreover, the regularization path is independent of the optimization algorithm, and thus easier to study. Here are some examples where $\bar{w}(B)/B$ converges.

A classical example is that if the loss has an exponential tail, then the regularization path converges to the maximum-margin direction (see [46], for the case of ℓ_1 regularization). It turns out that this is also true if the loss has a polynomial tail.

Proposition 2.9 ([22], Proposition 11)**.** If for some $a, b > 0$,

$$\lim_{z\to\infty} \frac{-\ell'(z)}{az^{-b}} = 1, \tag{2.41}$$

then $\lim_{B\to\infty} \bar{w}(B)/B$ exists.

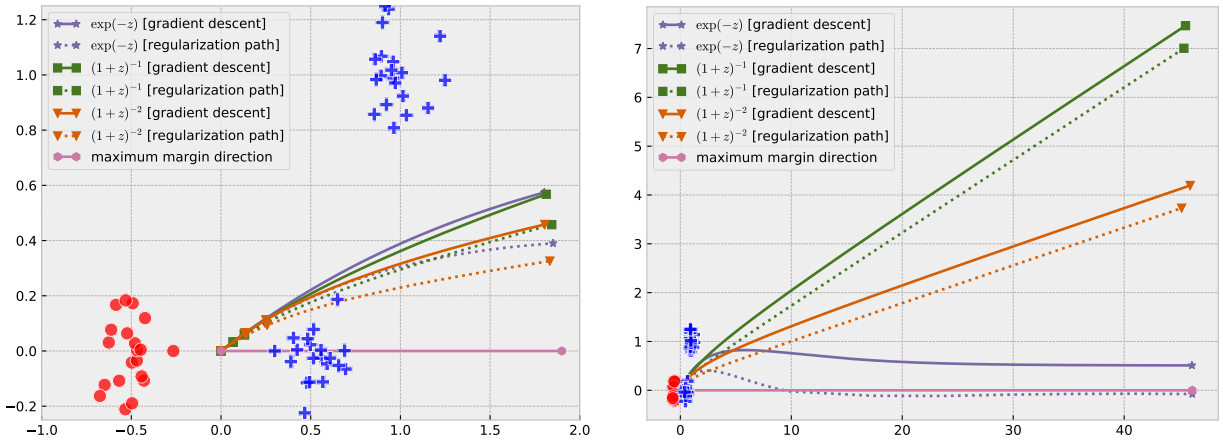
Moreover, it is indeed possible for a polynomial-tailed loss to induce a sub-optimal margin.

Proposition 2.10 ([22], Proposition 12)**.** For any $b > 0$, consider a loss function ℓ which equals z^{-b} for $z \geq 1$. There exists a dataset on which the maximum margin is a universal constant, while the regularization path with ℓ converges to a direction which has margin $\Theta(n^{-1/(b+1)})$.

Finally, Figure 2.1 gives some empirical results.

2.5 ADDITIONAL RELATED WORK

The first concrete studies showing an implicit bias of descent methods were for the ℓ_1-regularized case. Coordinate descent, when paired with the exponential loss, is implicitly



(a) Zoomed in. (b) Zoomed out.

Figure 2.1: Behavior of gradient descent and regularization path for three losses: the exponential loss $\exp(-z)$, and two polynomially-tailed losses $(1+z)^{-1}$ and $(1+z)^{-2}$ (with a quadratic extension along $z < 0$ for smoothness). The data has one negative (red) point cloud, and two positive (blue) point clouds; the upper positive cloud pulls the predictors trained with polynomially-tailed losses away from the maximum-margin direction, which points straight to the right.

biased towards ℓ_1-regularized solutions. This observation is the result of separate lines of work on descent methods and on regularization methods. On one hand, AdaBoost was shown to exhibit *positive margins*, meaning its predictions are not only correct, but in a certain sense robust [47]; indeed, with some further care on the descent step sizes, AdaBoost finds maximum-margin solutions [18, 48]. On the other hand, the ℓ_1-regularized solutions also converge to maximum-margin solutions as regularization strength is taken to 0 [46, 49].

On the implicit bias, [50] characterized the implicit bias of many other algorithms, such as mirror descent and natural gradient descent. [51] proved margin maximization for SGD. [52] also considered the relationship between GD iterates and regularization path, however they either focus on strongly convex objectives or margin-maximizing losses such as the exponential loss. [53, 54] analyzed the implicit bias in the adversarial training setting. [55] showed that momentum SGD and Adam also maximize the margin.

Chapter 3: Linear classifiers with general data

In the previous chapter, we focus on linear separable data. In this chapter, we will consider general data that may not be linearly separable.

First, in the distributional setting, we will consider the *agnostic learning* problem, where the optimal linear classifier can achieve a zero-one risk of $\mathrm{OPT} > 0$, and our goal is to compete with OPT. Previously, for a certain broad class of "well-behaved" distributions, [24] proved an $\widetilde{\Omega}(\mathrm{OPT})$ lower bound, while [25] proved an $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound. In Section 3.1, we close this gap by constructing a well-behaved distribution on which the global minimizer of the logistic risk only achieves $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ zero-one risk (cf. Theorem 3.1), matching the upper bound in [25]. On the other hand, we also show that we can overcome the $\Omega(\sqrt{\mathrm{OPT}})$ lower bound for logistic loss simply at the cost of one additional convex optimization step with the ReLU loss and attain $\widetilde{O}(\mathrm{OPT})$ zero-one risk (cf. Theorem 3.2). This two-step convex optimization algorithm is simpler than previous methods obtaining this guarantee, all of which require solving $O\left(\log(1/\mathrm{OPT})\right)$ minimization problems.

Second, we characterize the implicit bias of GD for general training data in Section 3.2. We show that the GD iterates are biased to follow a unique ray defined by the data: on the one hand, the direction of this ray is the maximum-margin predictor of a maximal linearly-separable subset of the data, and GD iterates converge to this ray in direction. On the other hand, this ray does not pass through the origin in general, and its offset is the bounded global optimum of the risk over the remaining data, and GD recovers this offset. This decomposition of training set is described formally in Theorem 3.3, while the implicit bias result is stated in Theorem 3.5.

3.1 AGNOSTIC LEARNING WITH THE LOGISTIC AND RELU LOSS

As before, we assume there is an unknown distribution P over $\mathbb{R}^d \times \{-1, +1\}$ to which we have access in the form of independent and identically distributed samples drawn from P. Our goal is to compete with a linear classifier $\bar{u}$ that achieves the optimal zero-one risk of $\mathrm{OPT} > 0$ over P. Alternatively, we can think that the labels of the examples are first generated by $\bar{u}$, and then an OPT fraction of the labels are adversarially corrupted.

A very natural heuristic for solving the problem is to use logistic regression. However, the analysis of logistic regression for this problem is still largely incomplete, even though it is one of the most fundamental algorithms in machine learning. One reason for this is that it can return extremely poor solutions in the worst case: [56] showed that the minimizer of logistic

risk may attain a zero-one risk of $1 - \mathrm{OPT}$ on an adversarially-constructed distribution.

As a result, much attention has been devoted to certain "well-behaved" distributions (see Assumption 3.1 for a subset of these conditions), for which much better results can be obtained. However, even under the "well-behaved" conditions, for any convex, nonincreasing, and nonconstant loss function, [24] proved an $\Omega\left(\mathrm{OPT}\ln(1/\mathrm{OPT})\right)$ lower bound for log-concave marginals and an $\Omega\left(\mathrm{OPT}^{1-1/s}\right)$ lower bound for s-heavy-tailed marginals. On the positive side, [25] assumed P_x satisfies a "soft-margin" condition: for anti-concentrated marginals such as isotropic log-concave marginals, this assumes $\Pr\left(\left|\langle \bar{u}, x\rangle\right| \leq \gamma\right) = O(\gamma)$ for any $\gamma > 0$. For sub-exponential distributions with soft-margins, they proved an $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound for gradient descent on the logistic loss, which can be improved to $O\left(\sqrt{\mathrm{OPT}}\right)$ for bounded distributions. Note that these upper bounds and the lower bounds in [24] do not match: if P_x is sub-exponential, then [24] only gave an $\widetilde{\Omega}(\mathrm{OPT})$ lower bound, while if P_x is s-heavy-tailed, then the upper bound in [25] becomes worse.

Here we first construct a distribution Q over $\mathbb{R}^2 \times \{-1, +1\}$, and prove a lower bound for logistic regression that matches the upper bound in [25], thereby closing the gap in [24, 25] (cf. Theorem 3.1). On the other hand, we describe a simple two-phase algorithm that achieves $\widetilde{O}\left(\mathrm{OPT}\right)$ error for general well-behaved distributions (cf. Theorem 3.2). Our two-phase algorithm involves logistic regression followed by SGD with the ReLU loss (i.e., the perceptron algorithm) with a restricted domain and warm start. Our algorithm is simple and sample-efficient: the output is guaranteed to have $O\left(\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon\right)$ zero-one risk using only $\widetilde{O}(d/\epsilon^2)$ samples. By contrast, [24] designed a nonconvex algorithm that achieves $O(\mathrm{OPT} + \epsilon)$ risk, but they need $\widetilde{O}(d/\epsilon^4)$ samples; other prior algorithms achieving $O(\mathrm{OPT} + \epsilon)$ error also involve solving multiple rounds of convex loss minimization [23, 57].

The contents in this section is based on [26]. In this paper, we also proved an $\widetilde{O}(\mathrm{OPT})$ upper bound for logistic regression under a radially-Lipschitz condition; it is omitted here.

Additional related work. The problem of agnostic learning of halfspaces has a long and rich history [58]. Here we survey the results most relevant to our work. It is well known that in the distribution independent setting, even *weak* agnostic learning is computationally hard [59, 60, 61]. As a result most algorithmic results have been obtained under assumptions on the marginal distribution P_x over the examples.

The work of [62] designed algorithms that achieve $\mathrm{OPT} + \epsilon$ error for any $\epsilon > 0$ in time $d^{\mathrm{poly}(\frac{1}{\epsilon})}$ for isotropic log-concave densities and for the uniform distribution over the hypercube. There is also recent evidence that removing the exponential dependence on $1/\epsilon$, even for Gaussian marginals is computationally hard [63, 64, 65].

As a result, another line of work aims to design algorithms with polynomial running time

and sample complexity (in d and $\frac{1}{\epsilon}$) and achieve an error of $g(\mathrm{OPT}) + \epsilon$, for g being a simple function. Along these lines [66] designed a polynomial-time algorithm that attains $\widetilde{O}(\mathrm{OPT}^{1/3}) + \epsilon$ zero-one risk for isotropic log-concave distributions. [23] improved the upper bound to $O(\mathrm{OPT}) + \epsilon$, using a localization-based algorithm. [67] further extended the algorithm to more general s-concave distributions. The work of [57] provided a *PTAS* guarantee: an error of $(1 + \eta)\mathrm{OPT} + \epsilon$ for any desired constant $\eta > 0$ via an improper learner.

Previously, [68] showed that stochastic gradient descent on a two-layer leaky ReLU network of any width achieves $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ zero-one risk, where OPT still denotes the best zero-one risk of a linear classifier. On the other hand, [27] showed that a wide two-layer ReLU network can even achieve the optimal Bayes risk, but their required width depends on a complexity measure that may be exponentially large in the worst case. It is an interesting open problem to see if a network with a reasonable width always reach a zero-one risk of $O(\mathrm{OPT})$.

Additional notation. Given $r > 0$, let $\mathcal{B}(r) := \left\{x \middle| \|x\|_2 \leq r\right\}$ denote the Euclidean ball with radius r. Given two nonzero vectors u and v, let $\varphi(u, v) \in [0, \pi]$ denote the angle between them.

Given a data distribution P over $\mathbb{R}^d \times \{-1, +1\}$, let P_x denote the marginal distribution of P on the feature space $\mathbb{R}^d$. We will frequently need the projection of the input features onto a two-dimensional subspace V; in such cases, it will be convenient to use polar coordinates (r, θ) for the associated calculations, such as parameterizing the density with respect to the Lebesgue measure as $p_V(r, \theta)$.

To be precise, given a nonincreasing loss function $\ell : \mathbb{R} \to \mathbb{R}$, let $\mathcal{R}_\ell$ and $\widehat{\mathcal{R}}_\ell$ denote the corresponding population and empirical risk. We will focus on the logistic loss the ReLU loss. For the logistic loss, let $\mathcal{R}_{\log} := \mathcal{R}_{\ell_{\log}}$ and $\widehat{\mathcal{R}}_{\log} := \widehat{\mathcal{R}}_{\ell_{\log}}$ for simplicity; for the ReLU loss, let $\mathcal{R}_{\mathrm{r}} := \mathcal{R}_{\ell_{\mathrm{r}}}$ and $\widehat{\mathcal{R}}_{\mathrm{r}} := \widehat{\mathcal{R}}_{\ell_{\mathrm{r}}}$ similarly. Recall that $\mathcal{R}_{0-1}(w) := \Pr_{(x,y)\sim P}\left(y \neq \mathrm{sign}\left(\langle w, x\rangle\right)\right)$ denote the population zero-one risk.

3.1.1 An $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ lower bound for logistic loss

In this subsection, we construct a distribution Q over $\mathbb{R}^2 \times \{-1, +1\}$ which satisfies standard regularity conditions in [24, 25], but the global minimizer w^* of the population logistic risk $\mathcal{R}_{\log}$ on Q only achieves a zero-one risk of $\Omega\left(\sqrt{\mathrm{OPT}}\right)$. Our focus on the global logistic optimizer is motivated by the lower bounds from [24]; in particular, this means that the large classification error is not caused by the sampling error.

The distribution Q has four parts Q_1, Q_2, Q_3, and Q_4, as described below. It can be verified that if $\mathrm{OPT} \leq 1/16$, the construction is valid.

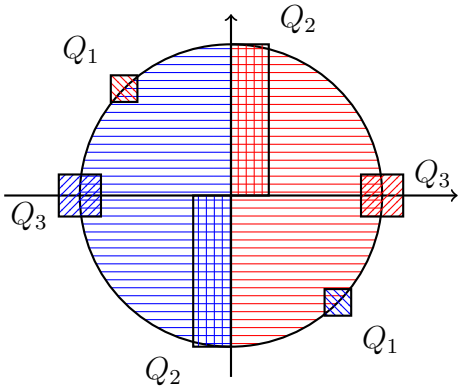


Figure 3.1: An illustration of Q when $\mathrm{OPT} = 1/16$. Red areas denote positive examples, while blue areas denote negative examples. The parts Q_1, Q_2 and Q_3 are marked in the figure, while Q_4 is supported on the unit circle and marked by horizontal lines.

1. The feature distribution of Q_1 consists of two squares: one has edge length $\sqrt{\frac{\mathrm{OPT}}{2}}$, center $\left(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right)$ and density 1, with label -1; the other has edge length $\sqrt{\frac{\mathrm{OPT}}{2}}$, center $\left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$, density 1, with label $+1$.

2. The feature distribution of Q_2 is supported on

$$\left(\left[0, \sqrt{\mathrm{OPT}}\right] \times [0,1]\right) \cup \left(\left[-\sqrt{\mathrm{OPT}}, 0\right] \times [-1,0]\right)$$

   with density 1, and the label is given by $\mathrm{sign}(x_1)$.

3. Let $q_3 := \frac{2}{3}\sqrt{\mathrm{OPT}}(1 - \mathrm{OPT})$, then Q_3 consists of two squares: one has edge length $\sqrt{\frac{q_3}{2}}$, center $(1,0)$, density 1 and label $+1$, and the other has edge length $\sqrt{\frac{q_3}{2}}$, center $(-1,0)$, density 1 and label -1.

4. The feature distribution of Q_4 is the uniform distribution over the unit ball $\mathcal{B}(1) := \left\{x \middle| \|x\| \le 1\right\}$ with density $q_4 := \frac{1-\mathrm{OPT}-2\sqrt{\mathrm{OPT}}-q_3}{\pi}$, and the label is given by $\mathrm{sign}(x_1)$.

Note that the correct label is given by $\mathrm{sign}(x_1)$ on Q_2, Q_3 and Q_4; therefore $\bar{u} := (1,0)$ is our ground-truth solution that is only wrong on Q_1. Here is our lower bound result.

Theorem 3.1. Suppose $\mathrm{OPT} \le 1/100$, and let Q_x denote the marginal distribution of Q on the feature space. It holds that $\mathbb{E}_{x\sim Q_x}[x] = 0$, and $\mathbb{E}_{x\sim Q_x}[x_1 x_2] = 0$, and $\mathbb{E}_{x\sim Q_x}[x_1^2 - x_2^2] = 0$. Moreover, the population logistic risk $\mathcal{R}_{\log}$ has a global minimizer w^*, and

$$\mathcal{R}_{0-1}(w^*) = \Pr\left(y \ne \mathrm{sign}\left(\langle w^*, x\rangle\right)\right) \ge \frac{\sqrt{\mathrm{OPT}}}{60\pi}.$$

We can normalize Q_x to unit variance and make it isotropic. Then it is easy to verify Q_x satisfies the "well-behaved" conditions in [24], and the "soft-margin" and "sub-exponential" conditions in [25]. Particularly, our lower bound matches the upper bound in [25].

Next we prove Theorem 3.1. First, we bound the density and support of Q_x.

Lemma 3.1. If $\mathrm{OPT} \le \frac{1}{100}$, then it holds that $q_3 \le \frac{1}{15}$, and $\frac{1}{2\pi} \le q_4 \le \frac{1}{\pi}$. As a result, Q_x is supported on $\mathcal{B}(2) := \{x \big| \|x\|_2 \le 2\}$ with its density bounded by 2.

Proof. For q_3, we have

$$q_3 = \frac{2}{3}\sqrt{\mathrm{OPT}}(1-\mathrm{OPT}) \le \frac{2}{3}\sqrt{\mathrm{OPT}} \le \frac{2}{3}\frac{1}{10} = \frac{1}{15}.$$

For Q_4, its total measure can be bounded as below:

$$1 - \mathrm{OPT} - 2\sqrt{\mathrm{OPT}} - q_3 \ge 1 - \frac{1}{100} - \frac{2}{10} - \frac{1}{15} \ge \frac{1}{2},$$

therefore $q_4 \ge \frac{1}{2\pi}$. The upper bound $q_4 \le \frac{1}{\pi}$ is trivial.

On the support of Q_x, note that for Q_1, the largest ℓ_2 norm is given by

$$1 + \frac{\sqrt{2}}{2}\sqrt{\frac{\mathrm{OPT}}{2}} \le 1 + \frac{1}{20} \le 2.$$

For Q_2, the largest ℓ_2 norm can be bounded by

$$1 + \sqrt{\mathrm{OPT}} \le 1 + \frac{1}{10} \le 2.$$

For Q_3, the largest ℓ_2 norm can be bounded by

$$1 + \frac{\sqrt{2}}{2}\sqrt{\frac{q_3}{2}} \le 1 + \frac{1}{2}\sqrt{\frac{1}{15}} \le 2.$$

Finally, it is easy to verify that if $\mathrm{OPT} \le \frac{1}{100}$, then Q_1, Q_2 and Q_3 do not overlap, therefore the density of Q is bounded by $1 + \frac{1}{\pi} \le 2$. QED.

The following fact is need in the proof of isotropy; its proof is straightforward and omitted.

Lemma 3.2. It holds that

$$\int_{a-\frac{\delta}{2}}^{a+\frac{\delta}{2}} \int_{b-\frac{\delta}{2}}^{b+\frac{\delta}{2}} xy \,\mathrm{d}y \,\mathrm{d}x = ab\delta^2, \quad \text{and} \quad \int_{a-\frac{\delta}{2}}^{a+\frac{\delta}{2}} \int_{b-\frac{\delta}{2}}^{b+\frac{\delta}{2}} (x^2 - y^2) \,\mathrm{d}y \,\mathrm{d}x = (a^2 - b^2)\delta^2.$$

Then we can prove that Q is isotropic up to a multiplicative factor.

Lemma 3.3. It holds that $\mathbb{E}_{x\sim Q_x}[x] = 0$, and $\mathbb{E}_{x\sim Q_x}[x_1 x_2] = 0$, and $\mathbb{E}_{x\sim Q_x}\left[x_1^2 - x_2^2\right] = 0$.

Proof. It follows from the symmetry of Q that $\mathbb{E}_{x\sim Q_x}[x] = 0$.

To verify $\mathbb{E}_{x\sim Q_x}[x_1 x_2] = 0$, note that the expectation of $x_1 x_2$ is 0 on Q_3 and Q_4, and thus we only need to check Q_1 and Q_2. First, due to Lemma 3.2, we have

$$\mathbb{E}_{(x,y)\sim Q_1}\left[x_1 x_2\right] = -\frac{\mathrm{OPT}}{2}.$$

Additionally,

$$\mathbb{E}_{(x,y)\sim Q_2}\left[x_1 x_2\right] = 2\int_0^{\sqrt{\mathrm{OPT}}}\int_0^1 x_1 x_2 \,\mathrm{d}x_2\,\mathrm{d}x_1 = \frac{\mathrm{OPT}}{2}.$$

Therefore $\mathbb{E}_{x\sim Q_x}[x_1 x_2] = 0$.

Finally, note that the expectation of $x_1^2 - x_2^2$ is 0 on Q_1 due to Lemma 3.2, and also 0 on Q_4 due to symmetry; therefore we only need to consider Q_2 and Q_3. We have

$$\mathbb{E}_{(x,y)\sim Q_2}\left[x_1^2 - x_2^2\right] = 2\int_0^{\sqrt{\mathrm{OPT}}}\int_0^1 \left(x_1^2 - x_2^2\right)\mathrm{d}x_2\,\mathrm{d}x_1 = \frac{2}{3}\mathrm{OPT}^{3/2} - \frac{2}{3}\sqrt{\mathrm{OPT}} = -q_3.$$

Since $\mathbb{E}_{(x,y)\sim Q_3}\left[x_1^2 - x_2^2\right] = q_3$ by Lemma 3.2, it follows that $\mathbb{E}_{x\sim Q_x}\left[x_1^2 - x_2^2\right] = 0$. QED.

Next we prove the risk lower bound. We only need to show that $\varphi(\bar{u}, w^*)$, the angle between $\bar{u}$ and w^*, is $\Omega\left(\sqrt{\mathrm{OPT}}\right)$, since it then follows that w^* is wrong on an $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ fraction of Q_4, which is enough since Q_4 accounts for more than half of the distribution Q. The first step is to show that by moving along the direction of $\bar{u}$ by a distance of $\Theta\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$, we can achieve a logistic risk of $O\left(\sqrt{\mathrm{OPT}}\right)$. For simplicity, in the remaining part of this subsection, we let $\mathcal{R}$ denote $\mathcal{R}_{\log}$. For $i = 1, 2, 3, 4$, we also let $\mathcal{R}_i(w) := \mathbb{E}_{(x,y)\sim Q_i}\left[\ell_{\log}\left(y\langle w, x\rangle\right)\right]$, and thereby $\mathcal{R}(w) := \sum_{i=1}^4 \mathcal{R}_i(w)$.

Lemma 3.4. Suppose $\mathrm{OPT} \le 1/100$, let $\bar{w} := (\bar{r}, 0)$ where $\bar{r} = \frac{3}{\sqrt{\mathrm{OPT}}}$, then $\mathcal{R}_{\log}(\bar{w}) \le 5\sqrt{\mathrm{OPT}}$.

Proof. We consider $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$ and $\mathcal{R}_4$ respectively.

1. For Q_1, note that the minimum of $y\langle \bar{w}, x\rangle$ is

$$-\left(\frac{\sqrt{2}}{2} + \frac{1}{2}\sqrt{\frac{\mathrm{OPT}}{2}}\right)\bar{r} = -\frac{3\sqrt{2}}{2}\frac{1}{\sqrt{\mathrm{OPT}}} - \frac{3\sqrt{2}}{4}.$$

Because $\ell_{\log}(z) \le -z + 1$ when $z \le 0$, and $\mathrm{OPT} \le \frac{1}{100}$, we have

$$\begin{aligned}\mathcal{R}_1(\bar{w}) \le \ell_{\log}\left(-\frac{3\sqrt{2}}{2}\frac{1}{\sqrt{\mathrm{OPT}}} - \frac{3\sqrt{2}}{4}\right)\cdot \mathrm{OPT} &\le \frac{3\sqrt{2}}{2}\sqrt{\mathrm{OPT}} + \left(\frac{3\sqrt{2}}{4}+1\right)\mathrm{OPT} \\ &\le \frac{3\sqrt{2}}{2}\sqrt{\mathrm{OPT}} + \left(\frac{3\sqrt{2}}{4}+1\right)\frac{1}{10}\sqrt{\mathrm{OPT}} \\ &\le \frac{5\sqrt{\mathrm{OPT}}}{2}.\end{aligned}$$

2. For Q_2, we have

$$\begin{aligned}\mathcal{R}_2(\bar{w}) = 2\int_0^{\sqrt{\mathrm{OPT}}}\int_0^1 \ell_{\log}(x_1\bar{r})\,\mathrm{d}x_2\,\mathrm{d}x_1 &= 2\int_0^{\sqrt{\mathrm{OPT}}} \ell_{\log}(x_1\bar{r})\,\mathrm{d}x_1 \\ &\le 2\int_0^{\sqrt{\mathrm{OPT}}} \exp(-x_1\bar{r})\,\mathrm{d}x_1 \\ &= \frac{2}{\bar{r}}\left(1-\exp\left(-\bar{r}\sqrt{\mathrm{OPT}}\right)\right) \le \frac{2}{\bar{r}},\end{aligned}$$

where we use $\ell_{\log}(z) \le \exp(-z)$.

3. For Q_3, the minimum of $y\langle \bar{w}, x\rangle$ is

$$\left(1 - \frac{1}{2}\sqrt{\frac{q_3}{2}}\right)\bar{r} \ge \frac{2\bar{r}}{3},$$

where we use $q_3 \le \frac{1}{15}$ by Lemma 3.1. Further note that $\ell_{\log}(z) \le 1/z$ when $z > 0$, we have

$$\mathcal{R}_3(\bar{w}) \le q_3\ell_{\log}\left(\frac{2\bar{r}}{3}\right) \le \frac{1/15}{2\bar{r}/3} \le \frac{1}{10\bar{r}}.$$

4. For Q_4,

$$\mathcal{R}_4(\bar{w}) = \int_0^1\int_0^{2\pi} \ell_{\log}\left(r\bar{r}\left|\cos(\theta)\right|\right) q_4 r\,\mathrm{d}\theta\,\mathrm{d}r \le \frac{1}{\pi}\int_0^1\int_0^{2\pi} \ell_{\log}\left(r\bar{r}\left|\cos(\theta)\right|\right) r\,\mathrm{d}\theta\,\mathrm{d}r,$$

where we use $q_4 \le \frac{1}{\pi}$ from Lemma 3.1. Lemma A.1 then implies

$$\mathcal{R}_4(\bar{w}) \le \frac{1}{\pi}\int_0^1 \frac{8\sqrt{2}}{\bar{r}}\,\mathrm{d}r = \frac{8\sqrt{2}}{\pi\bar{r}}.$$

Putting everything together, we have

$$\begin{aligned}\mathcal{R}(\bar{w}) &= \mathcal{R}_1(\bar{w}) + \mathcal{R}_2(\bar{w}) + \mathcal{R}_3(\bar{w}) + \mathcal{R}_4(\bar{w}) \\ &\le \frac{5\sqrt{\mathrm{OPT}}}{2} + \frac{2}{\bar{r}} + \frac{1}{10\bar{r}} + \frac{8\sqrt{2}}{\pi\bar{r}} \\ &\le \frac{5\sqrt{\mathrm{OPT}}}{2} + \frac{6}{\bar{r}} \le 5\sqrt{\mathrm{OPT}}.\end{aligned}$$

QED.

Next we consider the global minimizer w^* of $\mathcal{R}_{\log}$, which exists since $\mathcal{R}_{\log}$ has bounded sub-level sets. Let (r^*, θ^*) denote the polar coordinates of w^*. The plan is to assume $\theta^* \in \left[-\frac{\sqrt{\mathrm{OPT}}}{30}, \frac{\sqrt{\mathrm{OPT}}}{30}\right]$, and derive a contradiction, which would finish the proof.

In our construction, Q_3 and Q_4 are symmetric with respect to the horizontal axis, and they will induce the ground-truth solution. However, Q_1 and Q_2 are skew, and they will pull w^* above, meaning we actually have $\theta^* \in \left[0, \frac{\sqrt{\mathrm{OPT}}}{30}\right]$. The first observation is an upper bound on r^*: if r^* is too large, then the risk of w^* over Q_1 will already be larger than $\mathcal{R}_{\log}(\bar{w})$ for $\bar{w}$ constructed in Lemma 3.4, a contradiction.

Lemma 3.5. Suppose $\mathrm{OPT} \le 1/100$ and $\theta^* \in \left[0, \frac{\sqrt{\mathrm{OPT}}}{30}\right]$, then $r^* \le \frac{10}{\sqrt{\mathrm{OPT}}}$.

Proof. Let

$$u := \left(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right), \quad \text{and} \quad v := \left(\frac{\sqrt{2}}{2} - \frac{1}{2}\sqrt{\frac{\mathrm{OPT}}{2}}, -\frac{\sqrt{2}}{2} - \frac{1}{2}\sqrt{\frac{\mathrm{OPT}}{2}}\right).$$

Let ϕ denote the angle between u and v, then

$$\phi \le \tan(\phi) = \frac{\sqrt{2}}{2}\sqrt{\frac{\mathrm{OPT}}{2}} = \frac{\sqrt{\mathrm{OPT}}}{2} \le \frac{1}{20} \le \frac{\pi}{24},$$

and it follows that the angle between v and w^* is bounded by

$$\frac{\pi}{24} + \frac{\pi}{4} + \frac{\sqrt{\mathrm{OPT}}}{30} \le \frac{\pi}{24} + \frac{\pi}{4} + \frac{\pi}{24} = \frac{\pi}{3}.$$

Moreover, note that the maximum of $y\langle w^*, x\rangle$ on Q_1 is given by

$$-\langle w^*, v\rangle \le -r^*\|v\| \cos\left(\frac{\pi}{3}\right) \le -r^* \cos\left(\frac{\pi}{3}\right) = -\frac{r^*}{2}.$$

Additionally because $\ell_{\log}(z) > -z$, we have

$$\mathcal{R}(w^*) \geq \mathcal{R}_1(w^*) \geq \ell_{\log}\left(-\frac{r^*}{2}\right) \cdot \mathrm{OPT} > \frac{r^*}{2} \cdot \mathrm{OPT}.$$

If $r^* > \frac{10}{\sqrt{\mathrm{OPT}}}$, then $\mathcal{R}(w^*) > 5\sqrt{\mathrm{OPT}}$, which contradicts the definition of w^* in light of Lemma 3.4. Therefore $r^* \leq \frac{10}{\sqrt{\mathrm{OPT}}}$. QED.

However, our next lemma shows that under the above conditions, the gradient of $\mathcal{R}_{\log}$ at w^* does not vanish, which contradicts the definition of w^*.

Lemma 3.6. Suppose $\mathrm{OPT} \leq 1/100$, then for any $w = (r, \theta)$ with $0 \leq r \leq \frac{10}{\sqrt{\mathrm{OPT}}}$ and $0 \leq \theta \leq \frac{\sqrt{\mathrm{OPT}}}{30}$, it holds that $\nabla \mathcal{R}_{\log}(w) \neq 0$.

Proof. Let $w = (r, \theta)$, where $0 \leq r \leq \frac{10}{\sqrt{\mathrm{OPT}}}$ and $0 \leq \theta \leq \frac{\sqrt{\mathrm{OPT}}}{30}$. We will consider the projection of $\nabla \mathcal{R}(w)$ onto the direction $e_2 := (0, 1)$, and show this projection cannot be 0.

1. For Q_1, the gradient of this part has a negative inner product with e_2, due to the construction of Q_1 and the fact $\ell'_{\log} < 0$.

2. For Q_2, the inner product between e_2 and the gradient of this part is given by

$$2\int_0^{\sqrt{\mathrm{OPT}}} \int_0^1 \ell'_{\log}(x_1 w_1 + x_2 w_2) x_2 \, \mathrm{d}x_2 \, \mathrm{d}x_1. \tag{3.1}$$

   Note that $x_1 w_1 \leq r x_1$, while

$$x_2 w_2 = x_2 r \sin(\theta) \leq r\theta \leq \frac{10}{\sqrt{\mathrm{OPT}}} \frac{\sqrt{\mathrm{OPT}}}{30} = \frac{1}{3},$$

   and that $\ell'_{\log}$ is increasing, therefore

$$\ell'_{\log}(x_1 w_1 + x_2 w_2) \leq \ell'_{\log}\left(r x_1 + 1/3\right).$$

   We can then upper bound eq. (3.1) as follows:

$$\begin{aligned}
\text{eq. (3.1)} &\leq 2\int_0^{\sqrt{\mathrm{OPT}}} \int_0^1 \ell'_{\log}\left(r x_1 + \frac{1}{3}\right) x_2 \, \mathrm{d}x_2 \, \mathrm{d}x_1 \\
&= \int_0^{\sqrt{\mathrm{OPT}}} \ell'_{\log}\left(r x_1 + \frac{1}{3}\right) \mathrm{d}x_1 \\
&= \frac{1}{r}\left(\ell_{\log}\left(\frac{1}{3} + r\sqrt{\mathrm{OPT}}\right) - \ell_{\log}\left(\frac{1}{3}\right)\right).
\end{aligned}$$

Now we consider two cases. If $r\sqrt{\mathrm{OPT}} \le 2$, then it follows from the convexity of $\ell_{\log}$ that

$$\text{eq. (3.1)} \le \frac{1}{r}\ell'_{\log}\left(\frac{1}{3} + r\sqrt{\mathrm{OPT}}\right) r\sqrt{\mathrm{OPT}} \le \ell'_{\log}(3)\sqrt{\mathrm{OPT}} \le -\frac{\sqrt{\mathrm{OPT}}}{30}.$$

On the other hand, if $r\sqrt{\mathrm{OPT}} \ge 2$, then

$$\text{eq. (3.1)} \le \frac{1}{r}\left(\ell_{\log}\left(\frac{7}{3}\right) - \ell_{\log}\left(\frac{1}{3}\right)\right) \le \frac{\sqrt{\mathrm{OPT}}}{10}\left(\ell_{\log}\left(\frac{7}{3}\right) - \ell_{\log}\left(\frac{1}{3}\right)\right) \le -\frac{\sqrt{\mathrm{OPT}}}{30}.$$

Therefore, it always holds that eq. (3.1) $\le -\frac{\sqrt{\mathrm{OPT}}}{30}$.

3. For Q_3, the gradient of this part can have a positive inner product with e_2. For simplicity, let $\rho := \frac{1}{2}\sqrt{\frac{q_3}{2}}$. To upper bound this inner product, it is enough to consider the region given by

$$\left([1-\rho, 1+\rho] \times [-\rho, 0]\right) \cup \left([-1-\rho, -1+\rho] \times [0, \rho]\right).$$

Moreover, note that $y\langle w, x\rangle \ge 0$ on Q_3, therefore $\ell'_{\log}\left(y\langle w, x\rangle\right) \ge -\frac{1}{2}$. Therefore the inner product between e_2 and the gradient of Q_3 can be upper bounded by (note that $x_2 \le 0$ in the integral)

$$2\int_{1-\rho}^{1+\rho}\int_{-\rho}^{0} -\frac{1}{2}x_2 \,\mathrm{d}x_2 \,\mathrm{d}x_1 = \rho^3 = \frac{\sqrt{q_3}}{16\sqrt{2}}q_3 \le \frac{\sqrt{1/15}}{16\sqrt{2}}\frac{2}{3}\sqrt{\mathrm{OPT}} < \frac{\sqrt{\mathrm{OPT}}}{60}.$$

where we use $q_3 \le \frac{1}{15}$ by Lemma 3.1 and $q_3 \le \frac{2}{3}\sqrt{\mathrm{OPT}}$ by its definition.

4. For Q_4, we further consider two cases.

   (a) Consider the part of Q_4 with polar angles in $(-\frac{\pi}{2} + 2\theta, \frac{\pi}{2}) \cup (\frac{\pi}{2} + 2\theta, \frac{3\pi}{2})$. By symmetry, the gradient of this part is along the direction with polar angle $\pi + \theta$, and it has a negative inner product with e_2.

   (b) Consider the part of Q_4 with polar angles in $(-\frac{\pi}{2}, -\frac{\pi}{2} + 2\theta) \cup (\frac{\pi}{2}, \frac{\pi}{2} + 2\theta)$. We can verify that the gradient of this part has a positive inner product with e_2; moreover, since $-1 < \ell'_{\log} < 0$, this inner product can be upper bounded by

$$2\int_0^1\int_0^{2\theta} r'\cos(\theta')q_4 r' \,\mathrm{d}\theta' \,\mathrm{d}r' = 2q_4 \cdot \frac{1}{3} \cdot \sin(2\theta) \le \frac{4\theta}{3\pi} \le \frac{4}{3\pi}\frac{\sqrt{\mathrm{OPT}}}{30} < \frac{\sqrt{\mathrm{OPT}}}{60},$$

where we also use $q_4 \leq \frac{1}{\pi}$ and $\sin(z) \leq z$ for $z \geq 0$.

As a result, item 3 and item 4(b) cannot cancel item 2, and thus $\nabla \mathcal{R}(w)$ cannot be 0. QED.

Now we are ready to prove the risk lower bound of Theorem 3.1.

Proof of Theorem 3.1 risk lower bound. It is clear that $\mathcal{R}$ has bounded sub-level sets, and therefore can be globally minimized. Let the polar coordinates of the global minimizer be given by (r^*, θ^*), where $|\theta^*| \leq \pi$. Assume that $\theta^* \in \left[-\frac{\sqrt{\mathrm{OPT}}}{30}, \frac{\sqrt{\mathrm{OPT}}}{30}\right]$; due to Q_1 and Q_2, it actually follows that $\theta^* \in \left[0, \frac{\sqrt{\mathrm{OPT}}}{30}\right]$. Lemma 3.5 then implies $r^* \leq \frac{10}{\sqrt{\mathrm{OPT}}}$, and then Lemma 3.6 implies $\nabla \mathcal{R}(w^*) \neq 0$, a contradiction.

It then follows that w^* is wrong on a $\frac{\theta^*}{\pi}$ portion of Q_4. Since the total measure of Q_4 is more than half due to Lemma 3.1, we have

$$\mathcal{R}_{0-1}(w^*) \geq \frac{1}{2}\frac{\theta^*}{\pi} \geq \frac{\sqrt{\mathrm{OPT}}}{60\pi}.$$

QED.

3.1.2 A general framework for upper bound analysis

In this subsection, we present a general framework for analyzing the zero-one risk upper bound. In [26], we use this framework to show an $\widetilde{O}(\mathrm{OPT})$ upper bound for logistic regression under a radially-Lipschitz condition; for simplicity, we do not include this analysis here. However, this analysis framework is more versatile than that, as it can also recover the $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound for general well-behaved distributions with the logistic loss, and it can also handle the ReLU loss; we will use these results in the analysis of our two-phase algorithm.

First, we introduce some standard assumptions on the marginal distribution P_x. Because of the lower bound for s-heavy-tailed distributions from [24], to get an $\widetilde{O}(\mathrm{OPT})$ zero-one risk, we need to assume P_x has a light tail. Following [25], we will either consider a bounded distribution, or assume P_x is sub-exponential as defined below (cf. [69, Proposition 2.7.1 and Section 3.4.4]).

Definition 3.1. We say P_x is (α_1, α_2) sub-exponential for constants $\alpha_1, \alpha_2 > 0$, if for any unit vector v and any $t > 0$,

$$\Pr_{x \sim P_x}\left(\left|\langle v, x\rangle\right| \geq t\right) \leq \alpha_1 \exp\left(-t/\alpha_2\right).$$

We also need the next assumption, which is part of the "well-behaved" conditions from [24].

Assumption 3.1. There exist constants $U, R > 0$ and a function $\sigma : \mathbb{R}_+ \to \mathbb{R}_+$, such that if we project P_x onto an arbitrary two-dimensional subspace V, the corresponding density p_V satisfies $p_V(r,\theta) \geq 1/U$ for all $r \leq R$, and $p_V(r,\theta) \leq \sigma(r)$ for all $r \geq 0$, and $\int_0^\infty \sigma(r)\,\mathrm{d}r \leq U$, and $\int_0^\infty r\sigma(r)\,\mathrm{d}r \leq U$.

Note that for a broad class of distributions including isotropic log-concave distributions, the sub-exponential condition and Assumption 3.1 hold with α_1, α_2, U, R all being universal constants.

Next we describe our analysis framework. We will consider certain $\hat{w}$, such that $\mathcal{R}_\ell(\hat{w}) - \mathcal{R}_\ell(\bar{w})$ is small, where $\bar{w} := \|\hat{w}\|_2 \bar{u}$ and $\ell \in \{\ell_{\log}, \ell_{\mathrm{r}}\}$. For some concrete examples, $\hat{w}$ can be the global minimizer of $\mathcal{R}_\ell$, or some GD/SGD iterates whose excess risk is small. Our goal is to show that $\mathcal{R}_{0-1}(\hat{w})$ is also small.

The first step is to express $\mathcal{R}_\ell(\hat{w}) - \mathcal{R}_\ell(\bar{w})$ as the sum of three terms, and then bound them separately. The first term is given by

$$\mathcal{R}_\ell(\hat{w}) - \mathcal{R}_\ell(\bar{w}) - \mathbb{E}\left[\ell\left(\operatorname{sign}\left(\langle \bar{w}, x\rangle\right)\langle \hat{w}, x\rangle\right) - \ell\left(\operatorname{sign}\left(\langle \bar{w}, x\rangle\right)\langle \bar{w}, x\rangle\right)\right], \tag{3.2}$$

the second term is given by

$$\mathbb{E}\left[\ell\left(\operatorname{sign}\left(\langle \bar{w}, x\rangle\right)\langle \hat{w}, x\rangle\right) - \ell\left(\operatorname{sign}\left(\langle \hat{w}, x\rangle\right)\langle \hat{w}, x\rangle\right)\right], \tag{3.3}$$

and the third term is given by

$$\mathbb{E}\left[\ell\left(\operatorname{sign}\left(\langle \hat{w}, x\rangle\right)\langle \hat{w}, x\rangle\right) - \ell\left(\operatorname{sign}\left(\langle \bar{w}, x\rangle\right)\langle \bar{w}, x\rangle\right)\right], \tag{3.4}$$

where the expectations are taken over P_x.

We first bound term (3.2), which is the approximation error of replacing the true label y with the label given by $\bar{u}$. Since $\ell(-z) - \ell(z) = z$ for the logistic loss and ReLU loss, it follows that

$$\text{term (3.2)} = \mathbb{E}\left[\mathbb{1}_{y \neq \operatorname{sign}\left(\langle \bar{w}, x\rangle\right)} \cdot y\langle \bar{w} - \hat{w}, x\rangle\right].$$

The approximation error can be bounded as below, using the tail bound on P_x and the fact $\mathcal{R}_{0-1}(\bar{w}) = \mathrm{OPT}$.

Lemma 3.7. For $\ell \in \{\ell_{\log}, \ell_{\mathrm{r}}\}$, if $\|x\|_2 \le B$ almost surely,

$$\big|\text{term (3.2)}\big| \le B\|\bar{w} - \hat{w}\|_2 \cdot \mathrm{OPT}.$$

If P_x is (α_1, α_2)-sub-exponential, then

$$\big|\text{term (3.2)}\big| \le (1 + 2\alpha_1)\alpha_2\|\bar{w} - \hat{w}\|_2 \cdot \mathrm{OPT} \cdot \ln(1/\mathrm{OPT}).$$

Proof. Note that for both the logistic loss and the ReLU loss, it holds that $\ell(-z) - \ell(z) = z$, therefore

$$\text{term (3.2)} = \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y \neq \mathrm{sign}\left(\langle \bar{w}, x\rangle\right)} \cdot y\langle \bar{w} - \hat{w}, x\rangle\right], \tag{3.5}$$

It then follows from the triangle inequality that

$$\big|\text{term (3.2)}\big| \le \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y \neq \mathrm{sign}\left(\langle \bar{w}, x\rangle\right)}\big|\langle \bar{w} - \hat{w}, x\rangle\big|\right]$$

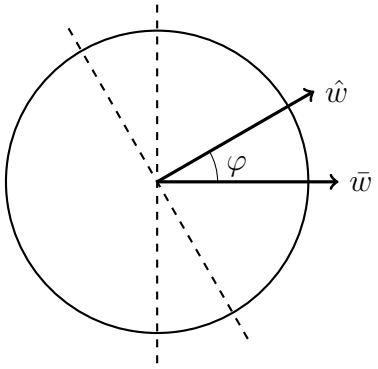
Now we can invoke Lemma A.2 with $w = \bar{w}$ and $w' = \bar{w} - \hat{w}$ to prove Lemma 3.7. QED.

Next we bound term (3.3).

Lemma 3.8. Under Assumption 3.1, for $\ell \in \{\ell_{\log}, \ell_{\mathrm{r}}\}$,

$$\text{term (3.3)} \ge \frac{4R^3}{3U\pi^2}\|\hat{w}\|_2\varphi(\hat{w}, \bar{w})^2.$$

Proof of Lemma 3.8. Note that in term (3.3), we only care about $\langle \hat{w}, x\rangle$ and $\langle \bar{w}, x\rangle$, therefore we can focus on the two-dimensional space spanned by $\bar{w}$ and $\hat{w}$. Let φ denote the angle between $\bar{w}$ and $\hat{w}$. Without loss of generality, we can consider the following graph, where we put $\bar{w}$ at angle 0, and $\hat{w}$ at angle φ.



We divide the graph into four parts given by different polar angles: (i) $(-\frac{\pi}{2}, -\frac{\pi}{2}+\varphi)$, (ii) $(-\frac{\pi}{2}+\varphi, \frac{\pi}{2})$, (iii) $(\frac{\pi}{2}, \frac{\pi}{2}+\varphi)$, and (iv) $(\frac{\pi}{2}+\varphi, \frac{3\pi}{2})$. Note that term (3.3) is 0 on parts (ii) and (iv), therefore we only need to consider parts (i) and (iii):

$$\begin{aligned}\text{term (3.3)} &= \mathbb{E}_{\text{(i) and (iii)}}\left[\ell\left(\operatorname{sign}\left(\langle \bar{w}, x\rangle\right)\langle \hat{w}, x\rangle\right) - \ell\left(\operatorname{sign}\left(\langle \hat{w}, x\rangle\right)\langle \hat{w}, x\rangle\right)\right] \\ &= \mathbb{E}_{\text{(i) and (iii)}}\left[-\operatorname{sign}\left(\langle \bar{w}, x\rangle\right)\langle \hat{w}, x\rangle\right].\end{aligned}$$

Here we use the fact that $\ell(-z) - \ell(z) = z$ for both the logistic loss and the ReLU loss.

For simplicity, let p denote the density of the projection of P_x onto the space spanned by $\hat{w}$ and $\bar{w}$. Under Assumption 3.1, we have

$$\begin{aligned}\text{term (3.3)} &= \mathbb{E}_{\text{(i) and (iii)}}\left[-\operatorname{sign}\left(\langle \bar{w}, x\rangle\right)\langle \hat{w}, x\rangle\right] \\ &= \int_0^\infty \int_{-\frac{\pi}{2}}^{-\frac{\pi}{2}+\varphi} -r\|\hat{w}\|_2 \cos(\varphi-\theta) p(r,\theta) r \,\mathrm{d}\theta \,\mathrm{d}r \\ &\quad + \int_0^\infty \int_{\frac{\pi}{2}}^{\frac{\pi}{2}+\varphi} r\|\hat{w}\|_2 \cos(\theta-\varphi) p(r,\theta) r \,\mathrm{d}\theta \,\mathrm{d}r \\ &\geq \frac{2}{U}\int_0^R \int_0^\varphi r\|\hat{w}\|_2 \sin(\theta) r \,\mathrm{d}\theta \,\mathrm{d}r \\ &= \frac{2R^3\|\hat{w}\|_2\left(1-\cos(\varphi)\right)}{3U} \geq \frac{4R^3\|\hat{w}\|_2\varphi^2}{3U\pi^2},\end{aligned}$$

where we use the fact that $1-\cos(\varphi) \geq \frac{2\varphi^2}{\pi^2}$ for all $\varphi \in [0, \pi]$. QED.

Lastly, we consider term (3.4). Note that it is 0 for the ReLU loss ℓ_{r}, because $\ell_{\mathrm{r}}(z) = 0$ when $z \geq 0$. For the logistic loss, term (3.4) can be bounded by $O(1/\|\hat{w}\|_2)$ in general as showed below, but we can also give a tighter bound with an additional radially-Lipschitz condition (see [26, Lemma 3.12] for more discussion).

Lemma 3.9. For $\ell = \ell_{\mathrm{r}}$, term (3.4) is 0. For $\ell = \ell_{\log}$, under Assumption 3.1,

$$\left|\text{term (3.4)}\right| \leq \frac{12U}{\|\hat{w}\|_2}.$$

Proof. For the ReLU loss, term (3.4) is 0 simply because $\ell_{\mathrm{r}}(z) = 0$ when $z \geq 0$.

For the logistic loss, we first show that under Assumption 3.1, for any $w \in \mathbb{R}^d$,

$$\mathbb{E}\left[\ell_{\log}\left(\left|\langle w, x\rangle\right|\right)\right] \leq \frac{12U}{\|w\|_2}.$$

Let v denote an arbitrary vector orthogonal to w, and let p denote the density of the projection of P_x onto the space spanned by w and v. Then we have

$$\mathbb{E}\left[\ell_{\log}\left(\left|\langle w, x\rangle\right|\right)\right] = \int_0^\infty \int_0^{2\pi} \ell_{\log}\left(r\|w\|_2\left|\cos(\theta)\right|\right) p(r,\theta) r \,\mathrm{d}\theta \,\mathrm{d}r.$$

Invoking Assumption 3.1, we have

$$\mathbb{E}\left[\ell_{\log}\left(\left|\langle w, x\rangle\right|\right)\right] \le \int_0^\infty \sigma(r) \left(\int_0^{2\pi} \ell_{\log}\left(r\|w\|_2\left|\cos(\theta)\right|\right) r \,\mathrm{d}\theta\right) \mathrm{d}r.$$

Lemma A.1 then implies

$$\mathbb{E}\left[\ell_{\log}\left(\left|\langle w, x\rangle\right|\right)\right] \le \int_0^\infty \sigma(r) \frac{8\sqrt{2}}{\|w\|_2} \,\mathrm{d}r.$$

Then it follows from Assumption 3.1 that

$$\mathbb{E}\left[\ell_{\log}\left(\left|\langle w, x\rangle\right|\right)\right] \le \frac{8\sqrt{2}U}{\|w\|_2} \le \frac{12U}{\|w\|_2}.$$

Now we have

$$\text{term (3.4)} \le \mathbb{E}\left[\ell_{\log}\left(\operatorname{sign}\left(\langle \hat{w}, x\rangle\right)\langle \hat{w}, x\rangle\right)\right] = \mathbb{E}\left[\ell_{\log}\left(\left|\langle \hat{w}, x\rangle\right|\right)\right] \le \frac{12U}{\|\hat{w}\|_2}.$$

Similarly, we can show

$$-\text{term (3.4)} \le \frac{12U}{\|\bar{w}\|_2} = \frac{12U}{\|\hat{w}\|_2}.$$

QED.

Lemmas 3.7 to 3.9 will be crucial in our analysis of the two-phase algorithm; they can also recover the $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound showed in [25]. Here we briefly discuss the ideas; some proof details are omitted, but they can be found in [26].

For simplicity, first let $\hat{w} = w^*$, the global optimizer of $\mathcal{R}_{\log}$, and assume $\|x\|_2 \le B$ almost surely. For simplicity, let φ denote $\varphi(\hat{w}, \bar{u})$. Lemmas 3.7 to 3.9 imply

$$C_1\|\hat{w}\|_2\varphi^2 \le B\|\bar{w} - \hat{w}\|_2 \cdot \mathrm{OPT} + \frac{C_2}{\|\hat{w}\|_2} \le B\|\hat{w}\|_2\varphi \cdot \mathrm{OPT} + \frac{C_2}{\|\hat{w}\|_2},$$

where $C_1 = 4R^3/(3U\pi^2)$ and $C_2 = 12U$. Now at least one of the following two cases is true:

1. $C_1\|\hat{w}\|_2\varphi^2 \le 3B\|\hat{w}\|_2\varphi \cdot \mathrm{OPT}$, which implies $\varphi = O(\mathrm{OPT})$;

2. $C_1\|\hat{w}\|_2\varphi^2 \le 3C_1/\|\hat{w}\|_2$, which implies $\varphi = O(1/\|\hat{w}\|_2)$.

Moreover, it follows from [26, Lemma 3.7] that $\|w^*\|_2 = \Omega\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$, and thus it always holds that $\varphi = O\left(\sqrt{\mathrm{OPT}}\right)$. Finally, we can get a zero-one risk bound by invoking the next result, which is basically [24, Claim 3.4].

Lemma 3.10. Under Assumption 3.1,

$$\mathcal{R}_{0-1}(\hat{w}) - \mathcal{R}_{0-1}(\bar{w}) \le \Pr\left(\mathrm{sign}\left(\langle \hat{w}, x\rangle\right) \ne \mathrm{sign}\left(\langle \bar{w}, x\rangle\right)\right) \le 2U\varphi(\hat{w}, \bar{w}).$$

Proof. Under Assumption 3.1, we have

$$\Pr\left(\mathrm{sign}\left(\langle \hat{w}, x\rangle\right) \ne \mathrm{sign}\left(\langle \bar{w}, x\rangle\right)\right) \le 2\varphi(\hat{w}, \bar{w}) \int_0^\infty \sigma(r) r \,\mathrm{d}r \le 2U\varphi(\hat{w}, \bar{w}).$$

QED.

Now for projected GD, we can obtain a similar guarantee as presented below. The proof is similar to the above analysis for w^*, but we also need to handle the optimization and generalization error of project GD; the proof details can be found on [26]. We present the bound in terms of the angle instead of zero-one risk for later application in the two-phase algorithm.

Lemma 3.11. Given the target error $\epsilon \in (0, 1)$ and the failure probability $\delta \in (0, 1/e)$, consider projected GD

$$w_{t+1} := \Pi_{\mathcal{B}(1/\sqrt{\epsilon})}\left[w_t - \eta \nabla \widehat{\mathcal{R}}_{\log}(w_t)\right].$$

If $\|x\|_2 \le B$ almost surely, then with $\eta = 4/B^2$, using $O\left(\frac{(B+1)^2 \ln(1/\delta)}{\epsilon^2}\right)$ samples and $O\left(\frac{B^2}{\epsilon^{3/2}}\right)$ iterations, with probability $1 - \delta$, projected GD outputs w_t with

$$\varphi(w_t, \bar{u}) = O\left(\sqrt{\mathrm{OPT} + \epsilon}\right).$$

If P_x is (α_1, α_2)-sub-exponential, then with $\eta = \widetilde{\Theta}(1/d)$, using $\widetilde{O}\left(\frac{d \ln(1/\delta)^3}{\epsilon^2}\right)$ samples and $\widetilde{O}\left(\frac{d \ln(1/\delta)^2}{\epsilon^{3/2}}\right)$ iterations, with probability $1 - \delta$, projected GD outputs w_t with

$$\varphi(w_t, \bar{u}) = O\left(\sqrt{\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon}\right).$$

3.1.3 An $\widetilde{O}(\mathrm{OPT})$ upper bound with ReLU loss

Note that in the analysis for the logistic loss presented in the previous subsection, we can only get an $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound because of term (3.4). However, as noted in Lemma 3.9, for the ReLU loss, term (3.4) is conveniently 0, which seems to suggest that we may overcome the $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ lower bound via the ReLU loss. However, note that we cannot simply optimize the ReLU risk $\mathcal{R}_{\mathrm{r}}$, since 0 is already a global minimizer of $\mathcal{R}_{\mathrm{r}}$. Instead, in our two-phase algorithm, we will first run logistic regression, and then run SGD with the ReLU loss on a restricted domain on which the norm is bounded below by 1.

Let

$$\mathcal{D} := \left\{ w \in \mathbb{R}^d \middle| \langle w, v \rangle \geq 1 \right\}, \tag{3.6}$$

the two-phase algorithm is described as below (the parameters η, T, etc. in this subsection are all chosen for the second phase):

1. Run projected GD under the settings of Lemma 3.11, and find a unit vector v such that $\varphi(v, \bar{u})$ is $O\left(\sqrt{\mathrm{OPT} + \epsilon}\right)$ for bounded distributions, or $O\left(\sqrt{\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon}\right)$ for sub-exponential distributions.

2. Run projected SGD over the domain $\mathcal{D}$ defined in eq. (3.6) starting from $w_0 := v$: at step t, we sample $(x_t, y_t) \sim P$, and let

$$w_{t+1} := \Pi_{\mathcal{D}} \left[w_t - \eta \ell_{\mathrm{r}}' \left(y_t \langle w_t, x_t \rangle \right) y_t x_t \right]. \tag{3.7}$$

   where we make the convention that $\ell_{\mathrm{r}}'(0) = -1$.

We show the following upper bound.

Theorem 3.2. Given the target error $\epsilon \in (0, 1/e)$, suppose Assumption 3.1 holds.

1. For bounded distributions, with $\eta = \Theta(\epsilon)$, for all $T = \Omega(1/\epsilon^2)$,

$$\mathbb{E}\left[\min_{0 \leq t < T} \mathcal{R}_{0-1}(w_t) \right] = O(\mathrm{OPT} + \epsilon).$$

2. For sub-exponential distributions, with $\eta = \Theta\left(\frac{\epsilon}{d \ln(d/\epsilon)^2} \right)$, for all $T = \Omega\left(\frac{d \ln(d/\epsilon)^2}{\epsilon^2} \right)$,

$$\mathbb{E}\left[\min_{0 \leq t < T} \mathcal{R}_{0-1}(w_t) \right] = O(\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon).$$

We prove Theorem 3.2 in the following. We will first handle the bounded case, and then consider the sub-exponential case.

Bounded distributions. Let $\bar{r} := 1/\langle v, \bar{u}\rangle$, and thus $\bar{r}\bar{u} \in \mathcal{D}$. At step t, we have

$$\|w_{t+1} - \bar{r}\bar{u}\|_2^2 \le \|w_t - \bar{r}\bar{u}\|_2^2 - 2\eta \left\langle \ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right) y_t x_t, w_t - \bar{r}\bar{u}\right\rangle + \eta^2 \ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right)^2 \|x_t\|_2^2. \tag{3.8}$$

Define

$$\mathcal{M}(w) := \mathbb{E}_{(x,y)\sim P}\left[-\ell_{\mathrm{r}}'\left(y\langle w, x\rangle\right)\right] = \mathcal{R}_{0-1}(w).$$

Taking expectation of eq. (3.8) w.r.t. (x_t, y_t), and note that $\|x\|_2 \le B$ almost surely and $(\ell_{\mathrm{r}}')^2 = -\ell_{\mathrm{r}}'$, we have

$$\begin{aligned}\mathbb{E}\left[\|w_{t+1} - \bar{r}\bar{u}\|_2^2\right] - \|w_t - \bar{r}\bar{u}\|_2^2 &\le -2\eta\left\langle \nabla\mathcal{R}_{\mathrm{r}}(w_t), w_t - \bar{r}\bar{u}\right\rangle + \eta^2 B^2 \mathcal{M}(w_t) \\ &\le -2\eta\left(\mathcal{R}_{\mathrm{r}}(w_t) - \mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u})\right) + \eta^2 B^2 \mathcal{M}(w_t). \end{aligned} \tag{3.9}$$

To continue, we first prove the following bound on $\mathcal{R}_{\mathrm{r}}(\bar{u})$.

Lemma 3.12. If $\|x\|_2 \le B$ almost surely, then $\mathcal{R}_{\mathrm{r}}(\bar{u}) \le B \cdot \mathrm{OPT}$, while if P_x is (α_1, α_2)-sub-exponential, then $\mathcal{R}_{\mathrm{r}}(\bar{u}) \le (1 + 2\alpha_1)\alpha_2 \cdot \mathrm{OPT} \cdot \ln(1/\mathrm{OPT})$.

Proof. Note that

$$\mathcal{R}_{\mathrm{r}}(\bar{u}) = \mathbb{E}_{(x,y)\sim P}\left[\ell_{\mathrm{r}}\left(y\langle \bar{u}, x\rangle\right)\right] = \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{\mathrm{sign}\left(\langle \bar{u}, x\rangle \neq y\right)} \left|\langle \bar{u}, x\rangle\right|\right].$$

It then follows from Lemma A.2 that if $\|x\|_2 \le B$ almost surely, then

$$\mathcal{R}_{\mathrm{r}}(\bar{u}) \le B \cdot \mathrm{OPT},$$

while if P_x is (α_1, α_2)-sub-exponential, then

$$\mathcal{R}_{\mathrm{r}}(\bar{u}) \le (1 + 2\alpha_1)\alpha_2 \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right).$$

QED.

Next we show the following key lemma, which follows from Lemmas 3.7 to 3.9, and the homogeneity of the ReLU loss ℓ_{r}.

Lemma 3.13. Suppose Assumption 3.1 holds. Consider an arbitrary $w \in \mathcal{D}$, and let φ

denote $\varphi(w, \bar{u})$. If $\|x\|_2 \le B$ almost surely, then

$$\mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u}) \le \mathcal{R}_{\mathrm{r}}(\|w\|_2\bar{u}) + O\left((\mathrm{OPT} + \epsilon)^2\right)$$

and

$$\mathcal{R}_{\mathrm{r}}(w) - \mathcal{R}_{\mathrm{r}}(\|w\|_2\bar{u}) \ge \frac{4R^3}{3U\pi^2}\|w\|_2\varphi^2 - B\|w\|_2\varphi \cdot \mathrm{OPT}.$$

If P_x is (α_1, α_2)-sub-exponential, then

$$\mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u}) \le \mathcal{R}_{\mathrm{r}}(\|w\|_2\bar{u}) + O\left(\left(\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon\right)^2\right)$$

and

$$\mathcal{R}_{\mathrm{r}}(w) - \mathcal{R}_{\mathrm{r}}(\|w\|_2\bar{u}) \ge \frac{4R^3}{3U\pi^2}\|w\|_2\varphi^2 - (1+2\alpha_1)\alpha_2\|w\|_2\varphi \cdot \mathrm{OPT} \cdot \ln(1/\mathrm{OPT}).$$

Proof. First assume $\|x\|_2 \le B$ almost surely. Note that ℓ_{r} is positive homogeneous, and thus for any positive constant c, we have $\mathcal{R}_{\mathrm{r}}(cw) = c\mathcal{R}_{\mathrm{r}}(w)$. Therefore, if $\bar{r} \le \|w\|_2$, then

$$\mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u}) = \frac{\bar{r}}{\|w\|_2}\mathcal{R}_{\mathrm{r}}(\|w\|_2\bar{u}) \le \mathcal{R}_{\mathrm{r}}(\|w\|_2\bar{u}).$$

If $\bar{r} \ge \|w\|_2$, then

$$\mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u}) = \mathcal{R}_{\mathrm{r}}\left(\|w\|_2\bar{u}\right) + \mathcal{R}_{\mathrm{r}}(\bar{u})\left(\bar{r} - \|w\|_2\right) \le \mathcal{R}_{\mathrm{r}}\left(\|w\|_2\bar{u}\right) + \mathcal{R}_{\mathrm{r}}(\bar{u})\left(\bar{r} - 1\right),$$

since $\|w\|_2 \ge 1$ for all $w \in \mathcal{D}$. Recall that

$$\bar{r} := \frac{1}{\langle v, \bar{u}\rangle} = \frac{1}{\cos\left(\varphi(v, \bar{u})\right)} \le \frac{1}{1 - \varphi(v, \bar{u})^2/2},$$

and therefore the first-phase of algorithm ensures $\bar{r} = 1 + O(\mathrm{OPT} + \epsilon)$ for bounded distributions, and $\bar{r} = 1 + O\left(\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon\right)$ for sub-exponential distributions. It then follows that for bounded distributions,

$$\begin{aligned}
\mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u}) &\le \mathcal{R}_{\mathrm{r}}\left(\|w_t\|_2\bar{u}\right) + \mathcal{R}_{\mathrm{r}}(\bar{u}) \cdot O(\mathrm{OPT} + \epsilon) \\
&\le \mathcal{R}_{\mathrm{r}}\left(\|w_t\|_2\bar{u}\right) + B \cdot \mathrm{OPT} \cdot O(\mathrm{OPT} + \epsilon) \\
&= \mathcal{R}_{\mathrm{r}}\left(\|w_t\|_2\bar{u}\right) + O\left((\mathrm{OPT} + \epsilon)^2\right),
\end{aligned}$$

where we apply Lemma 3.12 at the end. It also follows directly from Lemmas 3.7 to 3.9 that

$$\begin{aligned}\mathcal{R}_{\mathrm{r}}(w) - \mathcal{R}_{\mathrm{r}}(\|w\|_2\bar{u}) &\geq \frac{4R^3}{3U\pi^2}\|w\|_2\varphi^2 - B\left\|w - \|w\|_2\bar{u}\right\|_2 \cdot \mathrm{OPT} \\ &\geq \frac{4R^3}{3U\pi^2}\|w\|_2\varphi^2 - B\|w\|_2\varphi \cdot \mathrm{OPT}.\end{aligned}$$

The proof for the sub-exponential case is similar. QED.

Now we are ready to prove Theorem 3.2 for the bounded case.

Proof of Theorem 3.2, bounded distribution. Here we assume $\|x\|_2 \leq B$ almost surely. We will show that under the conditions of Theorem 3.2, then

$$\mathbb{E}\left[\min_{0\leq t<T}\varphi_t\right] = O(\mathrm{OPT} + \epsilon), \quad \text{where} \quad \varphi_t := \varphi(w_t, \bar{u}). \tag{3.10}$$

Further invoking Lemma 3.10 finishes the proof.

We first restate eq. (3.9): at step t, after taking the expectation with respect to (x_t, y_t), we have

$$\mathbb{E}\left[\|w_{t+1} - \bar{r}\bar{u}\|_2^2\right] \leq \|w_t - \bar{r}\bar{u}\|_2^2 - 2\eta\left(\mathcal{R}_{\mathrm{r}}(w_t) - \mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u})\right) + \eta^2 B^2 \mathcal{M}(w_t). \tag{3.11}$$

First, Lemma 3.13 implies

$$\begin{aligned}\mathcal{R}_{\mathrm{r}}(w_t) - \mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u}) &\geq \mathcal{R}_{\mathrm{r}}(w_t) - \mathcal{R}_{\mathrm{r}}(\|w_t\|_2\bar{u}) - O\left((\mathrm{OPT} + \epsilon)^2\right) \\ &\geq 2C_1\|w_t\|_2\varphi_t^2 - B\|w_t\|_2\varphi_t \cdot \mathrm{OPT} - O\left((\mathrm{OPT} + \epsilon)^2\right),\end{aligned}$$

where $C_1 := 2R^3/(3U\pi^2)$. Note that if $\varphi_t \leq B \cdot \mathrm{OPT}/C_1$, then eq. (3.10) holds; therefore in the following we assume

$$\varphi_t \geq \frac{B}{C_1} \cdot \mathrm{OPT}, \tag{3.12}$$

which implies

$$\mathcal{R}_{\mathrm{r}}(w_t) - \mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u}) \geq C_1\|w_t\|_2\varphi_t^2 - O\left((\mathrm{OPT} + \epsilon)^2\right) \geq C_1\varphi_t^2 - O\left((\mathrm{OPT} + \epsilon)^2\right), \tag{3.13}$$

since $\|w\|_2 \geq 1$ for all $w \in \mathcal{D}$.

On the other hand, eq. (3.12) and Lemma 3.10 imply

$$\mathcal{M}(w_t) = \mathcal{R}_{0-1}(w_t) \le \mathrm{OPT} + 2U\varphi_t \le \left(\frac{C_1}{B} + 2U\right)\varphi_t.$$

Let

$$C_2 := \frac{C_1}{\left(\frac{C_1}{B} + 2U\right)B^2}.$$

Note that if $\varphi_t \le \epsilon$, then eq. (3.10) is true; otherwise we can assume $\epsilon \le \varphi_t$, and let $\eta = C_2\epsilon$, we have

$$\eta B^2 \mathcal{M}(w_t) \le C_2\epsilon B^2 \left(\frac{C_1}{B} + 2U\right)\varphi_t = C_1\epsilon\varphi_t \le C_1\varphi_t^2. \tag{3.14}$$

Now eqs. (3.11), (3.13) and (3.14) imply

$$\begin{aligned}\mathbb{E}\left[\|w_{t+1} - \bar{r}\bar{u}\|_2^2\right] &\le \|w_t - \bar{r}\bar{u}\|_2^2 - 2\eta C_1\varphi_t^2 + \eta C_1\varphi_t^2 + \eta \cdot O\left((\mathrm{OPT} + \epsilon)^2\right) \\ &= \|w_t - \bar{r}\bar{u}\|_2^2 - \eta C_1\varphi_t^2 + \eta \cdot O\left((\mathrm{OPT} + \epsilon)^2\right).\end{aligned}$$

Taking the expectation and average, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t<T}\varphi_t^2\right] \le \frac{\|w_0 - \bar{r}\bar{u}\|_2^2}{\eta C_1 T} + \frac{O\left((\mathrm{OPT} + \epsilon)^2\right)}{C_1}.$$

Note that

$$\|w_0 - \bar{r}\bar{u}\|_2 = \tan(\varphi_0) = O\left(\sqrt{\mathrm{OPT} + \epsilon}\right),$$

and also recall $\eta = C_2\epsilon$, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t<T}\varphi_t^2\right] \le \frac{O(\mathrm{OPT} + \epsilon)}{C_1 C_2 \epsilon T} + \frac{O\left((\mathrm{OPT} + \epsilon)^2\right)}{C_1}.$$

Letting $T = \Omega(1/\epsilon^2)$, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t<T}\varphi_t^2\right] \le O\left((\mathrm{OPT} + \epsilon)\epsilon\right) + O\left((\mathrm{OPT} + \epsilon)^2\right) = O\left((\mathrm{OPT} + \epsilon)^2\right),$$

and thus eq. (3.10) holds. QED.

Sub-exponential distributions. Next we consider sub-exponential distributions. We first prove the following bound on the square of norm.

Lemma 3.14. Suppose P_x is (α_1, α_2)-sub-exponential. Given any threshold $\tau > 0$, it holds that

$$\mathbb{E}\left[\|x\|_2^2 \mathbb{1}_{\|x\|_2 \geq \tau}\right] \leq d\alpha_1 \left(\tau^2 + 2\sqrt{d}\alpha_2 \tau + 2d\alpha_2^2\right) \exp\left(-\frac{\tau}{\sqrt{d}\alpha_2}\right).$$

Proof. First recall that

$$\Pr\left(\|x\|_2 \geq \tau\right) \leq \sum_{j=1}^{d} \Pr\left(|x_j| \geq \frac{\tau}{\sqrt{d}}\right) \leq d\alpha_1 \exp\left(-\frac{\tau}{\sqrt{d}\alpha_2}\right) =: \delta(\tau).$$

Let $\mu(\tau) := \Pr\left(\|x\|_2 \geq \tau\right)$. Integration by parts gives

$$\mathbb{E}\left[\|x\|_2^2 \mathbb{1}_{\|x\|_2 \geq \tau}\right] = \int_\tau^\infty r^2 \cdot (-\,\mathrm{d}\mu(r)) = \tau^2 \mu(\tau) + \int_\tau^\infty 2r\mu(r)\,\mathrm{d}r \leq \tau^2 \delta(\tau) + \int_\tau^\infty 2r\delta(r)\,\mathrm{d}r.$$

Calculation gives

$$\mathbb{E}\left[\|x\|_2^2 \mathbb{1}_{\|x\|_2 \geq \tau}\right] \leq d\alpha_1 \left(\tau^2 + 2\sqrt{d}\alpha_2 \tau + 2d\alpha_2^2\right) \exp\left(-\frac{\tau}{\sqrt{d}\alpha_2}\right).$$

QED.

Now we are ready to prove Theorem 3.2 for sub-exponential distributions.

Proof of Theorem 3.2, sub-exponential distributions. At step t, we have

$$\begin{aligned}
\|w_{t+1} - \bar{r}\bar{u}\|_2^2 &\leq \|w_t - \bar{r}\bar{u}\|_2^2 - 2\eta \left\langle \ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right) y_t x_t, w_t - \bar{r}\bar{u} \right\rangle + \eta^2 \ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right)^2 \|x_t\|_2^2 \\
&= \|w_t - \bar{r}\bar{u}\|_2^2 - 2\eta \left\langle \ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right) y_t x_t, w_t - \bar{r}\bar{u} \right\rangle - \eta^2 \ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right) \|x_t\|_2^2,
\end{aligned} \tag{3.15}$$

where we use $(\ell_{\mathrm{r}}')^2 = -\ell_{\mathrm{r}}'$. Next we bound $\mathbb{E}_{(x_t, y_t)}\left[-\ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right) \|x_t\|_2^2\right]$.

Let $\tau := \sqrt{d}\alpha_2 \ln(d/\epsilon)$. When $\|x_t\|_2 \leq \tau$, we have

$$\mathbb{E}\left[-\ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right) \|x_t\|_2^2 \mathbb{1}_{\|x_t\|_2 \leq \tau}\right] \leq \tau^2 \mathcal{M}(w_t) \leq d\alpha_2^2 \mathcal{M}(w_t) \cdot \ln(d/\epsilon)^2.$$

On the other hand, when $\|x_t\|_2 \geq \tau$, Lemma 3.14 implies

$$\begin{aligned}\mathbb{E}\left[-\ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right)\|x_t\|_2^2 \mathbb{1}_{\|x_t\|_2\geq\tau}\right] &\leq \mathbb{E}\left[\|x_t\|_2^2 \mathbb{1}_{\|x_t\|_2\geq\tau}\right] \\ &\leq d\alpha_1 \cdot O\left(d\ln(d/\epsilon)^2\right)\cdot\frac{\epsilon}{d} \\ &= O\left(d\epsilon\ln(d/\epsilon)^2\right),\end{aligned}$$

where we also use $\ln(1/\epsilon) > 1$, since $\epsilon < 1/e$. To sum up,

$$\mathbb{E}_{(x_t,y_t)}\left[-\ell_{\mathrm{r}}'\left(y_t\langle w_t, x_t\rangle\right)\|x_t\|_2^2\right] \leq Cd\left(\mathcal{M}(w_t)+\epsilon\right)\cdot\ln(d/\epsilon)^2$$

for some constant C.

Now taking the expectation with respect to (x_t, y_t) on both sides of eq. (3.15), we have

$$\mathbb{E}\left[\|w_{t+1}-\bar{r}\bar{u}\|_2^2\right] \leq \|w_t-\bar{r}\bar{u}\|_2^2 - 2\eta\left(\mathcal{R}_{\mathrm{r}}(w_t)-\mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u})\right)+\eta^2 Cd\left(\mathcal{M}(w_t)+\epsilon\right)\cdot\ln(d/\epsilon)^2. \tag{3.16}$$

Similarly to the bounded case, we will show that

$$\mathbb{E}\left[\min_{0\leq t<T}\varphi_t\right] = O\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})+\epsilon\right), \quad \text{where} \quad \varphi_t := \varphi(w_t,\bar{u}). \tag{3.17}$$

First, Lemma 3.13 implies

$$\begin{aligned}\mathcal{R}_{\mathrm{r}}(w_t)-\mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u}) &\geq \mathcal{R}_{\mathrm{r}}(w_t)-\mathcal{R}_{\mathrm{r}}(\|w_t\|_2\bar{u}) - O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})+\epsilon\right)^2\right) \\ &\geq 2C_1\|w_t\|_2\varphi_t^2 - C_2\|w_t\|_2\varphi_t\cdot\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) \\ &\quad - O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})+\epsilon\right)^2\right),\end{aligned}$$

where $C_1 := 2R^3/(3U\pi^2)$ and $C_2 = (1+2\alpha_1)\alpha_2$. Note that if $\varphi_t \leq C_2\cdot\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})/C_1$, then eq. (3.17) holds; therefore in the following we assume

$$\varphi_t \geq \frac{C_2}{C_1}\cdot\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}), \tag{3.18}$$

which implies

$$\begin{aligned}\mathcal{R}_{\mathrm{r}}(w_t)-\mathcal{R}_{\mathrm{r}}(\bar{r}\bar{u}) &\geq C_1\|w_t\|_2\varphi_t^2 - O\left(\left(\mathrm{OPT}\cdot(1/\mathrm{OPT})+\epsilon\right)^2\right) \\ &\geq C_1\varphi_t^2 - O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})+\epsilon\right)^2\right),\end{aligned} \tag{3.19}$$

since $\|w\|_2 \geq 1$ for all $w \in \mathcal{D}$.

On the other hand, for $\mathrm{OPT} \leq 1/e$, eq. (3.18) and Lemma 3.10 imply

$$\mathcal{M}(w_t) = \mathcal{R}_{0-1}(w_t) \leq \mathrm{OPT} + 2U\varphi_t \leq \left(\frac{C_1}{C_2} + 2U\right)\varphi_t.$$

Let

$$C_2 := \frac{C_1}{\left(\frac{C_1}{C_2} + 2U\right)C}.$$

Note that if $\varphi_t \leq \epsilon$, then eq. (3.17) is true; otherwise we can assume $\epsilon \leq \varphi_t$, and let $\eta = \frac{C_2\epsilon}{d\ln(d/\epsilon)^2}$, we have

$$\begin{aligned}
\eta C d\left(\mathcal{M}(w_t) + \epsilon\right)\ln(d/\epsilon)^2 &= \frac{C_2\epsilon}{d\ln(d/\epsilon)^2} C d\mathcal{M}(w_t)\cdot\ln(d/\epsilon)^2 + \frac{C_2\epsilon}{d\ln(d/\epsilon)^2} C d\epsilon\cdot\ln(d/\epsilon)^2 \\
&\leq C_2\epsilon C\left(\frac{C_1}{C_2} + 2U\right)\varphi_t + C_2 C\epsilon^2 \\
&= C_1\epsilon\varphi_t + O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon\right)^2\right) \\
&\leq C_1\varphi_t^2 + O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon\right)^2\right). \qquad (3.20)
\end{aligned}$$

Now eqs. (3.16), (3.19) and (3.20) imply

$$\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - \bar{r}\bar{u}\|_2^2\right] &\leq \|w_t - \bar{r}\bar{u}\|_2^2 - 2\eta C_1\varphi_t^2 + \eta C_1\varphi_t^2 + \eta\cdot O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon\right)^2\right) \\
&= \|w_t - \bar{r}\bar{u}\|_2^2 - \eta C_1\varphi_t^2 + \eta\cdot O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon\right)^2\right).
\end{aligned}$$

Taking the expectation and average, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t<T}\varphi_t^2\right] \leq \frac{\|w_0 - \bar{r}\bar{u}\|_2^2}{\eta C_1 T} + \frac{O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon\right)^2\right)}{C_1}.$$

Note that $\|w_0 - \bar{r}\bar{u}\| = \tan(\varphi_0) = O\left(\sqrt{\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon}\right)$, and also recall $\eta = \frac{C_2\epsilon}{d\ln(d/\epsilon)^2}$, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t<T}\varphi_t^2\right] \leq \frac{O\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon\right)d\ln(d/\epsilon)^2}{C_1 C_2\epsilon T} + \frac{O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon\right)^2\right)}{C_1}.$$

Letting $T = \Omega\left(\frac{d\ln(d/\epsilon)^2}{\epsilon^2}\right)$, we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t<T}\varphi_t^2\right] &\leq O\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})+\epsilon\right)\cdot\epsilon + O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})+\epsilon\right)^2\right)\\ &= O\left(\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})+\epsilon\right)^2\right),\end{aligned}$$

and thus eq. (3.17) holds. QED.

3.2 CHARACTERIZATION OF THE IMPLICIT BIAS

In this section, we will characterize the implicit bias of GD on a general training set. Specifically, we will show that GD, when applied to the task of logistic regression, outputs iterates which are biased to follow a unique ray defined by the data. The direction of this ray is the maximum-margin predictor of a maximal linearly separable subset of the data; the GD iterates converge to this ray *in direction*. The ray does not pass through the origin in general, and its offset is the bounded global optimum of the risk over the remaining data; gradient descent recovers this offset at a rate $O((\ln t)^2/\sqrt{t})$.

This section is based on [17]. For simplicity, here we will often focus on the exponential loss; however, the logistic loss was also handled in [17].

Additional notation. Recall that the training set is denoted by $\{(x_i, y_i)\}_{i=1}^n$, and for simplicity we let $z_i := y_i x_i$. Assume $\|z_i\|_2 \leq 1$ without loss of generality, and collect them into a matrix $Z \in \mathbb{R}^{n\times d}$ whose i-th row is given by $z_i^\top$. Given a loss function ℓ, for any positive integer k and any $v \in \mathbb{R}^k$, let $L(v) := \sum_{i=1}^k \ell(v_i)$, whereby $\widehat{\mathcal{R}}(w) := L(Zw)/n$, with gradient $\nabla\widehat{\mathcal{R}}(w) := Z^\top \nabla L(Zw)/n$. Note that we allow L to have varying input dimension, which will be convenient in the following analysis.

As in Theorem 3.3, the matrix Z defines a unique division of $\mathbb{R}^d$ into a direct sum of subspaces $\mathbb{R}^d = S \oplus S^\perp$. The rows of Z are either within S or S^c (i.e., $\mathbb{R}^d \setminus S$), and without loss of generality reorder the examples (and permute the rows of Z) so that $Z = \begin{bmatrix} Z_c \\ Z_S \end{bmatrix}$ where the rows of Z_S are within S and the rows of Z_c are within S^c; tying this to the earlier discussion, Z_c is the maximal linearly separable part of the data, and Z_S consists of the remaining data. Furthermore, let Π_S and $\Pi_\perp$ respectively denote orthogonal projection onto S and $S^\perp$, and define $Z_\perp := \Pi_\perp Z_c$ for simplicity, where each row of Z_c is orthogonally projected onto $S^\perp$.

By this notation,

$$\Pi_\perp \nabla(L \circ Z)(w) = \Pi_\perp \begin{bmatrix} Z_c \\ Z_S \end{bmatrix}^\top \nabla L\left(\begin{bmatrix} Z_c \\ Z_S \end{bmatrix} w\right) = \Pi_\perp \begin{bmatrix} Z_c \\ Z_S \end{bmatrix}^\top \begin{bmatrix} \nabla L(Z_c w) \\ \nabla L(Z_S w) \end{bmatrix} = \begin{bmatrix} Z_\perp \\ 0 \end{bmatrix}^\top \begin{bmatrix} \nabla L(Z_c w) \\ \nabla L(Z_S w) \end{bmatrix}$$
$$= Z_\perp^\top \nabla L(Z_c w),$$

which has made use of L at varying input dimensions.

GD here starts with $w_0 := 0$, and thereafter set $w_{j+1} := w_j - \eta_j \nabla \widehat{\mathcal{R}}(w_j)$. It is convenient to define $\gamma_j := \left\| \nabla(\ln \widehat{\mathcal{R}})(w_j) \right\|_2 = \left\| \nabla \widehat{\mathcal{R}}(w_j) \right\|_2 / \widehat{\mathcal{R}}(w_j)$ and $\hat{\eta}_j := \eta_j \widehat{\mathcal{R}}(w_j)$, whereby

$$\|w_t\|_2 \leq \sum_{j<t} \eta_j \left\| \nabla \widehat{\mathcal{R}}(w_j) \right\|_2 = \sum_{j<t} \hat{\eta}_j \gamma_j.$$

3.2.1 Problem structure

In this subsection, we characterize the unique ray $\{\bar{v} + r \cdot \bar{u} : r \geq 0\}$ to which GD iterates are biased. To build towards this, first consider the following examples.

Linearly separable. Consider the data at right in Figure 3.2: a blue circle of positive points, and a red circle of negative points. The data is *linearly separable:* there exist vectors $u \in \mathbb{R}^d$ with positive margin, meaning $\min_i \langle u, x_i y_i \rangle > 0$. Taking any such u and extending it to infinity will achieve 0 risk, but this is not what gradient descent chooses. Constraining u to have unit norm, a unique maximum margin point $\bar{u} = -e_1$ is obtained. The green gradient descent iterates follow $\bar{u}$ exactly.

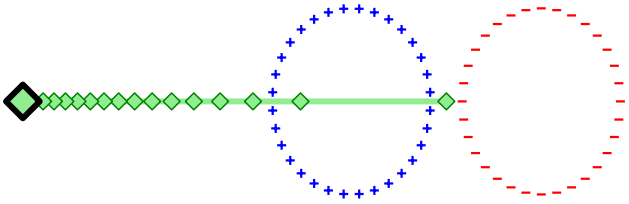


Figure 3.2: Separable.

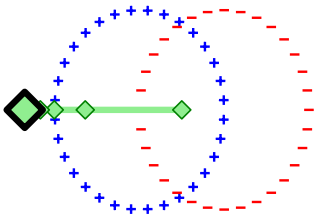


Figure 3.3: Strongly convex.

Strong convexity. Now consider moving the circles of data in Figure 3.2 until they overlap, obtaining Figure 3.3. This data is *not* linearly separable; indeed, given any nonzero vector $u \in \mathbb{R}^d$, there exist data points incorrectly classified by u, and therefore extending u indefinitely will cause the risk to also increase to infinity. It follows that the risk itself is 0-coercive [70], and moreover strongly convex over bounded subsets, with a unique optimum $\bar{v}$. Gradient descent converges towards $\bar{v}$.

An intermediate setting. The preceding two settings either had the circles overlapping, or far apart. What if they are pressed together so that they touch at the origin? Excluding the point at the origin, the circles may still be separated with the maximum margin separator $\bar{u} = -e_1$ from the linearly separable instance Figure 3.2. This example is our first taste of the general ray $\{\bar{v} + r \cdot \bar{u} : r \geq 0\}$, albeit still with some triviality: $\bar{v} = 0$. Specifically, $\bar{u}$ is the maximum margin separator of all data excluding the point at the origin; the risk in this instance is bounded below by $\ell(0)/n$, which is the necessary error on the point at the origin; the global optimum for that single point is $\bar{v} = 0$.

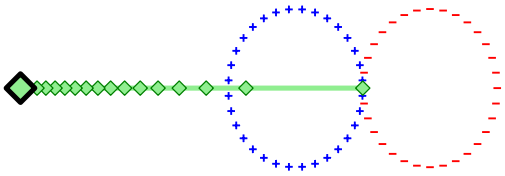


Figure 3.4: Mixed data.

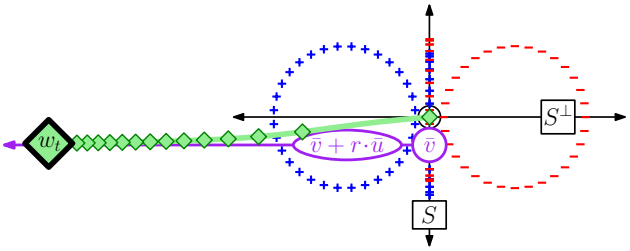


Figure 3.5: The general case.

The general case. Combining elements from the preceding examples, the general case may be characterized as follows; it appears in Figure 3.5, with all relevant objects labeled. In the general case, the dataset consists of a *maximal linearly separable subset* Z_c, with the remaining data falling into a subset Z_S over which the empirical risk is strongly convex. Specifically, Z_c is constructed with the following greedy procedure: for each example (x_i, y_i), include it in Z_c if there exists u_i with $\langle u_i, x_i y_i \rangle > 0$ and $\min_j \langle u_j, x_j y_j \rangle \geq 0$. The aggregate $u := \sum_{i \in Z_c} u_i$ satisfies $\langle u, x_i y_i \rangle > 0$ for $i \in Z_c$ and $\langle u, x_i y_i \rangle = 0$ otherwise. Therefore Z_c can be strictly separated by some vector u orthogonal to Z_S; let $\bar{u}$ denote the maximum margin separator of Z_c, which is also orthogonal to Z_S.

Turning now to Z_S, any vector v which is correct on some $(x_i, y_i) \in Z_S$ (i.e., $\langle u, x_i y_i \rangle > 0$) must also be incorrect on some other example (x_j, y_j) in Z_S (i.e., $\langle v, x_j y_j \rangle < 0$); otherwise, (x_i, y_i) would have been included in Z_c! Consequently, as in Figure 3.3 above, the empirical risk restricted to Z_S is strongly convex, with a unique optimum $\bar{v}$. The gradient descent iterates follow the ray $\{\bar{v} + r \cdot \bar{u} : r \geq 0\}$, which means they are globally optimal along Z_S, and achieve zero risk and follow the maximum margin direction $\bar{u}$.

Turning back to the construction in Figure 3.5, the linearly separable data Z_c is the two red and blue circles, while Z_S consists of data points on the vertical axis. The points in Z_S do not affect $\bar{u}$, and have been adjusted to move $\bar{v}$ away from 0, where it rested in Figure 3.4.

The above constructions are made rigorous in the following Theorem 3.3; its proof follows

the intuition above.

Theorem 3.3. The rows of Z can be uniquely partitioned into matrices (Z_S, Z_c), with a corresponding pair of orthogonal subspaces $(S, S^\perp)$ where $S = \operatorname{span}(Z_S^\top)$ satisfying the following properties.

1. **(Strongly convex part.)** If ℓ is twice continuously differentiable with $\ell'' > 0$, and $\ell \geq 0$, and $\lim_{z\to\infty} \ell(z) = 0$, then $L \circ Z$ is strongly convex over compact subsets of S, and $L \circ Z_S$ admits a unique minimizer $\bar{v}$ with $\inf_{w\in\mathbb{R}^d} L(Zw) = \inf_{v\in S} L(Z_S v) = L(Z_S \bar{v})$.

2. **(Separable part.)** If Z_c is nonempty (and thus so is $Z_\perp$), then $Z_\perp$ is linearly separable. The maximum margin is given by

$$\gamma := \max\left\{ \min_i (Z_\perp u)_i : \|u\|_2 = 1 \right\} = \min\left\{ \left\| Z_\perp^\top q \right\|_2 : q \geq 0, \sum_i q_i = 1 \right\} > 0,$$

   and the maximum margin solution $\bar{u}$ is the unique optimum to the primal problem, satisfying $\bar{u} = Z_\perp^\top \bar{q}/\gamma$ for every dual optimum $\bar{q}$. If $\ell \geq 0$ and $\lim_{z\to\infty} \ell(z) = 0$, then

$$\inf_{w\in\mathbb{R}^d} L(Zw) = L(Z_S\bar{v}) + \lim_{r\to\infty} L\left(Z_c(\bar{v} + r\bar{u})\right) = L(Z_S\bar{v}).$$

Before proving Theorem 3.3, note the following result characterizing margin maximization over $S^\perp$, which can be showed by applying Lemma 2.3 to $Z_\perp$.

Lemma 3.15. Suppose $Z_\perp$ has $n_c > 0$ rows and there exists u with $Z_\perp u > 0$. Then

$$\gamma := \max\left\{ \min_i (Z_\perp u)_i : \|u\|_2 = 1 \right\} = \min\left\{ \left\| Z_\perp^\top q \right\|_2 : q \geq 0, \sum_i q_i = 1 \right\} > 0.$$

Moreover there exists a unique nonzero primal optimum $\bar{u}$, and every dual optimum $\bar{q}$ satisfies $\bar{u} = Z_\perp^\top \bar{q}/\gamma$.

The proof of Theorem 3.3 follows.

Proof of Theorem 3.3. Partition the rows of Z into Z_c and Z_S as follows. For each row i, put it in Z_c if there exists u_i so that $Zu_i \geq 0$ (coordinate-wise) and $(Zu_i)_i > 0$; otherwise, when no such u_i exists, add this row to Z_S. Define $S := \operatorname{span}(Z_S^\top)$, the linear span of the rows of Z_S. This has the following consequences.

- To start, $S^\perp = \operatorname{span}(Z_S^\top)^\perp = \ker(Z_S) \supseteq \ker(Z)$.

- For each row i of Z_c, the corresponding u_i has $Z_S u_i = 0$, since otherwise $Zu_i \geq 0$ implies there would be a positive coordinate of $Z_S u_i$, and this row should be in Z_c not Z_S. Combining this with the preceding point, $u_i \in \ker(Z_S) = S^\perp$. Define $\tilde{u} := \sum_i u_i \in S^\perp$, whereby $Z_c\tilde{u} > 0$ and $Z_S\tilde{u} = 0$. Lastly, $\tilde{u} \in \ker(Z_S)$ implies moreover that $Z_\perp \tilde{u} = Z_c\tilde{u} > 0$. As such, when Z_c has a positive number of rows, Lemma 3.15 can be applied, resulting in the desired unique $\bar{u} = Z_\perp^\top \bar{q}/\gamma \in S^\perp$ with $\gamma > 0$.

- S, $S^\perp$, Z_S, Z_c, and $\bar{u}$ are unique and constructed from Z, with no dependence on ℓ.

- If Z_c is empty, the proof is trivial, thus suppose Z_c is nonempty. Since $\lim_{z\to\infty} \ell(z) = 0$,

$$0 \leq \inf_{w\in\mathbb{R}^d} L(Z_c w) \leq \inf_{w\in S^\perp} L(Z_c w) \leq \inf_{u\in S^\perp} L(Z_c u) \leq \lim_{r\to\infty} L(r \cdot Z_c \bar{u}) = 0.$$

  Since these inequalities start and end with 0, they are equalities, and consequently $\inf_{w\in\mathbb{R}^d} = \inf_{u\in S^\perp} L(Z_c u) = 0$. Moreover,

$$\begin{aligned}\inf_{w\in\mathbb{R}^d} L(Aw) &\leq \inf_{\substack{v\in S\\ u\in S^\perp}} \big(L(Z_S(u+v) + L(Z_c(u+v))\big) = \inf_{v\in S}\big(L(Z_S v) + \inf_{u\in S^\perp} L(Z_c(u+v))\big)\\ &\leq \left(\inf_{v\in S} L(Z_S v)\right) + \left(\inf_{u\in S^\perp} L(Z_c u)\right) = \left(\inf_{v\in S} L(Z_S v)\right) \leq \inf_{w\in\mathbb{R}^d} L(Aw).\end{aligned}$$

  which again is in fact a chain of equalities.

- For every $v \in S$ with $\|v\|_2 > 0$, there exists a row a of Z_S such that $\langle a, v\rangle < 0$. To see this, suppose contradictorily that $Z_S v \geq 0$. It cannot hold that $Z_S v = 0$, since $v \neq 0$ and $\ker(Z_S) \subseteq S^\perp$. this means $Z_S v \geq 0$ and moreover $(Z_S v)_i > 0$ for some i. But since $Z\bar{u} \geq 0$ and $Z_c\bar{u} > 0$, then for a sufficiently large $r > 0$, $Z(v + r\bar{u}) \geq 0$ and $(Z_S(v + r\bar{u}))_i > 0$, which means row i of Z_S should have been in Z_c, a contradiction.

- Consider any $v \in S \setminus \{0\}$. By the preceding point, there exists a row a of Z_S such that $\langle a, v\rangle < 0$. Since $\ell(0) > 0$ (because $\ell'' > 0$) and $\lim_{z\to\infty} \ell(z) = 0$, there exists $r > 0$ so that $\ell(-r\langle a, v\rangle) = \ell(0)/2$. By convexity, for any $t > 0$, setting $\alpha := r/(t + r)$ and noting $\alpha\langle a, tv\rangle + (1-\alpha)\langle a, -rv\rangle = 0$,

$$\alpha\ell(t\langle a, v\rangle) \geq \ell(0) - (1-\alpha)\ell(-r\langle a, v\rangle) = \left(\frac{1+\alpha}{2}\right)\ell(0),$$

  thus $\ell(t\langle a, v\rangle) \geq \left(\frac{1+\alpha}{2\alpha}\right)\ell(0) = \left(\frac{t+2r}{2r}\right)\ell(0)$, and

$$\lim_{t\to\infty} \frac{L(tAv) - L(0)}{t} \geq \lim_{t\to\infty} \frac{\ell(t\langle a, v\rangle) - n\ell(0)}{t} \geq \lim_{t\to\infty} \frac{\ell(0)}{2r}\left(\frac{(t+2r) - 2nr}{t}\right) > 0.$$

Consequently, $L \circ Z$ has compact sublevel sets over S [70, Proposition B.3.2.4].

- Note $\nabla^2 L(v) = \operatorname{diag}(\ell''(v_1), \dots, \ell''(v_n))$. Moreover, since $\ker(Z) \subseteq S^\perp$, then the image $B_0 := \{Zv : v \in S, \|v\|_2 = 1\}$ over the surface of the ball in S through Z is a collection of vectors with positive length. Thus for any compact subset $S_0 \subseteq S$,

$$\begin{aligned}\inf_{\substack{v_1 \in S_0 \\ v_2 \in S, \|v_2\|_2 = 1}} v_2^\top \nabla^2 (L \circ A)(v_1) v_2 &= \inf_{\substack{v_1 \in S_0 \\ v_2 \in S, \|v_2\|_2 = 1}} (Av_2)^\top \nabla^2 L(Av_1)(Av_2) \\ &= \inf_{\substack{v_1 \in S_0 \\ v_3 \in B_0}} v_3^\top \nabla^2 L(Av_1) v_3 \\ &\geq \inf_{\substack{v_1 \in S_0 \\ v_3 \in B_0}} \|v_3\|_2^2 \min_i \ell''((v_1)_i) > 0,\end{aligned}$$

  the final inequality since the minimization is of a continuous function over a compact set, thus attained at some point, and the infimand is positive over the domain. Consequently, $L \circ Z$ is strongly convex over compact subsets of S.

- Since $L \circ Z$ is strongly convex over S and moreover has bounded sublevel sets over S, it attains a unique optimum over S.

QED.

3.2.2 Risk convergence

In this subsection, we prove risk convergence guarantees for the exponential loss, which will also be useful in the characterization of the implicit bias.

Note that the exponential loss is not globally smooth, but we can still use local smoothness and prove the following result. The proof is based on the convergence guarantee for AdaBoost [12]. Recall the definitions $\gamma_j := \left\|\nabla(\ln \widehat{\mathcal{R}})(w_j)\right\|_2 = \left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2 / \widehat{\mathcal{R}}(w_j)$ and $\hat{\eta}_j := \eta_j \widehat{\mathcal{R}}(w_j)$.

Lemma 3.16. Suppose ℓ is convex, $|\ell'| \leq \ell$, $\ell'' \leq \ell$, and $\hat{\eta}_j = \eta_j \widehat{\mathcal{R}}(w_j) \leq 1$. Then

$$\widehat{\mathcal{R}}(w_{j+1}) \leq \widehat{\mathcal{R}}(w_j) - \eta_j \left(1 - \frac{\eta_j \widehat{\mathcal{R}}(w_j)}{2}\right) \left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2^2 = \widehat{\mathcal{R}}(w_j) \left(1 - \hat{\eta}_j (1 - \hat{\eta}_j/2) \gamma_j^2\right)$$

and thus

$$\widehat{\mathcal{R}}(w_t) \leq \widehat{\mathcal{R}}(w_0) \prod_{j<t} \left(1 - \hat{\eta}_j (1 - \hat{\eta}_j/2) \gamma_j^2\right) \leq \widehat{\mathcal{R}}(w_0) \exp\left(-\sum_{j<t} \hat{\eta}_j (1 - \hat{\eta}_j/2) \gamma_j^2\right).$$

Additionally, $\|w_t\|_2 \le \sum_{j<t} \hat{\eta}_j \gamma_j$.

Next we prove Lemma 3.16. For convenience, for the rest of this subsection define $w' := w - \eta \nabla \widehat{\mathcal{R}}(w) = w - \eta Z^\top \nabla L(Zw)/n$, and suppose throughout that ℓ is twice differentiable.

Lemma 3.17. For any $w \in \mathbb{R}^d$,

$$\widehat{\mathcal{R}}(w') \le \widehat{\mathcal{R}}(w) - \eta \left\| \nabla \widehat{\mathcal{R}}(w) \right\|_2^2 + \frac{\eta^2}{2} \left\| \nabla \widehat{\mathcal{R}}(w) \right\|_2^2 \max_{v \in [w,w']} \sum_i \ell''(Z_i v)/n.$$

Proof. By Taylor expansion,

$$\widehat{\mathcal{R}}(w') \le \widehat{\mathcal{R}}(w) - \eta \left\| \nabla \widehat{\mathcal{R}}(w) \right\|_2^2 + \frac{1}{2} \max_{v \in [w,w']} \sum_i (Z_i(w-w'))^2 \ell''(Z_i v)/n.$$

By Hölder's inequality,

$$\max_{v \in [w,w']} \sum_i (Z_i(w-w'))^2 \ell''(Z_i v) \le \max_{v \in [w,w']} \left\| Z(w-w') \right\|_\infty^2 \sum_i \ell''(Z_i v).$$

Since $\max_i \|Z_i\|_2 \le 1$,

$$\begin{aligned} \left\| Z(w-w') \right\|_\infty^2 = \eta^2 \left\| Z \nabla \widehat{\mathcal{R}}(w) \right\|_\infty^2 &= \eta^2 \max_i \left\langle Z_i, \nabla \widehat{\mathcal{R}}(w) \right\rangle^2 \\ &\le \eta^2 \max_i \|Z_i\|^2 \left\| \nabla \widehat{\mathcal{R}}(w) \right\|_2^2 \le \eta^2 \left\| \nabla \widehat{\mathcal{R}}(w) \right\|_2^2. \end{aligned}$$

Thus

$$\widehat{\mathcal{R}}(w') \le \widehat{\mathcal{R}}(w) - \eta \left\| \nabla \widehat{\mathcal{R}}(w) \right\|_2^2 + \frac{\eta^2}{2} \left\| \nabla \widehat{\mathcal{R}}(w) \right\|_2^2 \max_{v \in [w,w']} \sum_i \ell''(Z_i v)/n.$$

QED.

Lemma 3.18. Suppose $|\ell'|, \ell'' \le \ell$ and ℓ is convex. Then, for any $w \in \mathbb{R}^d$,

$$\max_{v \in [w,w']} \sum_i \ell''(Z_i v)/n \le \max \left\{ \widehat{\mathcal{R}}(w), \widehat{\mathcal{R}}(w') \right\}.$$

Define $\hat{\eta} := \eta \widehat{\mathcal{R}}(w)$ and suppose $\hat{\eta} \le 1$; then $\widehat{\mathcal{R}}(w') \le \widehat{\mathcal{R}}(w)$ and

$$\widehat{\mathcal{R}}(w') \le \widehat{\mathcal{R}}(w) \left(1 - \hat{\eta}(1 - \hat{\eta}/2) \frac{\left\| \nabla \widehat{\mathcal{R}}(w) \right\|_2^2}{\widehat{\mathcal{R}}(w)^2} \right).$$

Proof. Since $\ell'' \le \ell$ and ℓ is convex,

$$\max_{v\in[w,w']} \sum_i \ell''(Z_i v)/n \le \max_{v\in[w,w']} \sum_i \ell(Z_i v)/n = \max_{v\in[w,w']} \widehat{\mathcal{R}}(v) = \max\left\{\widehat{\mathcal{R}}(w), \widehat{\mathcal{R}}(w')\right\}.$$

Combining this, the choice of η, and Lemma 3.17,

$$\begin{aligned}\widehat{\mathcal{R}}(w') &\le \widehat{\mathcal{R}}(w) - \eta\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2^2 + \frac{\eta^2}{2}\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2^2 \max\left\{\widehat{\mathcal{R}}(w), \widehat{\mathcal{R}}(w')\right\} \\ &= \widehat{\mathcal{R}}(w) - \frac{\hat\eta\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2^2}{\widehat{\mathcal{R}}(w)}\left(1 - \frac{\hat\eta}{2}\frac{\max\left\{\widehat{\mathcal{R}}(w), \widehat{\mathcal{R}}(w')\right\}}{\widehat{\mathcal{R}}(w)}\right).\end{aligned}$$

Finally, suppose $\widehat{\mathcal{R}}(w') > \widehat{\mathcal{R}}(w)$; since $\hat\eta \le 1$ and $|\ell'| \le \ell$ and $\max_i \|Z_i\|_2 \le 1$,

$$\frac{\widehat{\mathcal{R}}(w')}{\widehat{\mathcal{R}}(w)} - 1 \le \frac{\hat\eta\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2^2}{\widehat{\mathcal{R}}(w)^2}\left(\frac{\hat\eta}{2}\frac{\widehat{\mathcal{R}}(w')}{\widehat{\mathcal{R}}(w)} - 1\right) \le \hat\eta\left(\frac{\hat\eta}{2}\frac{\widehat{\mathcal{R}}(w')}{\widehat{\mathcal{R}}(w)} - 1\right) \le \frac{1}{2}\frac{\widehat{\mathcal{R}}(w')}{\widehat{\mathcal{R}}(w)} - 1,$$

a contradiction. Therefore $\widehat{\mathcal{R}}(w') \le \widehat{\mathcal{R}}(w)$, which in turn implies

$$\widehat{\mathcal{R}}(w') \le \widehat{\mathcal{R}}(w) - \frac{\hat\eta\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2^2}{\widehat{\mathcal{R}}(w)}\left(1 - \frac{\hat\eta}{2}\right).$$

QED.

Together, these pieces prove the desired smoothness inequality.

Proof of Lemma 3.16. For any $j < t$, by Lemma 3.18 and the definition of γ_j,

$$\widehat{\mathcal{R}}(w_{j+1}) \le \widehat{\mathcal{R}}(w_j)\left(1 - \hat\eta_j(1 - \hat\eta_j/2)\frac{\left\|\nabla\widehat{\mathcal{R}}(w_j)\right\|_2^2}{\widehat{\mathcal{R}}(w_j)^2}\right) = \widehat{\mathcal{R}}(w_j)\left(1 - \hat\eta_j(1 - \hat\eta_j/2)\gamma_j^2\right).$$

Applying this recursively gives the bound.

Lastly,

$$\|w_t\|_2 = \left\|\sum_{j<t} \hat\eta_j \nabla\widehat{\mathcal{R}}(w_j)/\widehat{\mathcal{R}}(w_j)\right\|_2 \le \sum_{j<t}\left\|\hat\eta_j \nabla\widehat{\mathcal{R}}(w_j)/\widehat{\mathcal{R}}(w_j)\right\|_2 = \sum_{j<t}\hat\eta_j\gamma_j.$$

QED.

Now by combining Lemma 3.16 and Lemma 2.13, we can prove the following risk bound.

Theorem 3.4. For $\ell \in \{\ell_{\log}, \ell_{\exp}\}$, given step sizes $\eta_j \leq 1$ and $w_0 = 0$, then for any $t \geq 1$,

$$\widehat{\mathcal{R}}(w_t) - \inf_{w \in \mathbb{R}^d} \widehat{\mathcal{R}}(w) \leq \frac{\exp\left(\|\bar{v}\|_2\right)}{t} + \frac{\|\bar{v}\|_2^2 + \ln(t)^2/\gamma^2}{2\sum_{j=0}^{t-1} \eta_j}.$$

Proof. Note that both $\ell_{\exp}$ and $\ell_{\log}$ satisfy the conditions of Lemma 3.16. Moreover, since $\widehat{\mathcal{R}}(w_0) = \ell(0) \leq 1$, it follows that as long as $\eta_j \leq 1$, we have $\hat{\eta}_j = \eta_j \widehat{\mathcal{R}}(w_j) \leq 1$ and $\widehat{\mathcal{R}}(w_{j+1}) \leq \widehat{\mathcal{R}}(w_j)$. Therefore Lemma 3.16 implies

$$\widehat{\mathcal{R}}(w_{j+1}) \leq \widehat{\mathcal{R}}(w_j) - \eta_j \left(1 - \frac{\eta_j \widehat{\mathcal{R}}(w_j)}{2}\right) \left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2^2 \leq \widehat{\mathcal{R}}(w_j) - \frac{\eta_j}{2} \left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2^2,$$

and it then follows from Lemma 2.13 that for any $\bar{w} \in \mathbb{R}^d$,

$$\widehat{\mathcal{R}}(w_t) \leq \widehat{\mathcal{R}}(\bar{w}) + \frac{\|\bar{w}\|_2^2}{2\sum_{j<t} \eta_j}.$$

The proof is done by letting $\bar{w} := \bar{v} + \bar{u}\ln(t)/\gamma$, since

$$L(Z\bar{w}) = L(Z_S\bar{v}) + L(Z_c\bar{w}) \leq \inf_w L(Aw) + n\exp\left(\|\bar{v}\|_2 - \ln(t)\right) = \inf_w L(Aw) + \frac{n\exp\left(\|\bar{v}\|_2\right)}{t},$$

where we invoke Theorem 3.3 and use $\ell_{\log} \leq \ell_{\exp}$ and $\|z_i\|_2 \leq 1$. QED.

3.2.3 The implicit bias analysis

Here is our implicit bias result.

Theorem 3.5. Consider the exponential loss. Let the learning rate $\eta_j = 1/\sqrt{j+1}$, then it holds that

$$\|\Pi_S w_t\|_2 = \Theta(1) \quad \text{and} \quad \|\Pi_S w_t - \bar{v}\|_2^2 = O\left(\frac{\ln(t)^2}{\sqrt{t}}\right),$$

and if Z_c is nonempty, then $w_t/\|w_t\|_2 \to \bar{u}$.

Below we prove Theorem 3.5. Note that to illustrate the key proof ideas, we only focus on the exponential loss and asymptotic convergence here; [17] further handled the logistic loss and provided convergence rates.

The analysis over S is easy. For convenience, define

$$\widehat{\mathcal{R}}_S(w) := \frac{L(Z_S w)}{n}, \quad \widehat{\mathcal{R}}_c(w) := \frac{L(Z_c w)}{n}, \quad \bar{\mathcal{R}} := \inf_{w \in \mathbb{R}^d} \widehat{\mathcal{R}}(w),$$

and note $\widehat{\mathcal{R}}(w) = \widehat{\mathcal{R}}_S(w) + \widehat{\mathcal{R}}_c(w)$. Convergence over S is a consequence of strong convexity (cf. Theorem 3.3) and risk convergence (cf. Theorem 3.4).

Lemma 3.19. Let λ denote the modulus of strong convexity of $\widehat{\mathcal{R}}_S$ over the 1-sublevel set (guaranteed positive by Theorem 3.3). Then for any $t \geq 1$,

$$\|\Pi_S w_t - \bar{v}\|_2^2 \leq \frac{2}{\lambda} \min \left\{ 1, \ \frac{\exp\left(\|\bar{v}\|_2\right)}{t} + \frac{\|\bar{v}\|_2^2 + \ln(t)^2/\gamma^2}{2\sum_{j=0}^{t-1} \eta_j}. \right\}$$

Proof. By Theorem 3.3, $\widehat{\mathcal{R}}_S(\bar{v}) = \bar{\mathcal{R}}$. Thus, by strong convexity,

$$\|\Pi_S w_t - \bar{v}\|_2^2 \leq \frac{2}{\lambda} \left(\widehat{\mathcal{R}}_S(w) - \widehat{\mathcal{R}}_S(\bar{v}) \right) \leq \frac{2}{\lambda} \left(\widehat{\mathcal{R}}(w_t) - \inf_{w \in \mathbb{R}^d} \widehat{\mathcal{R}}(w) \right).$$

The bound follows by noting $\widehat{\mathcal{R}}(w_t) \leq \widehat{\mathcal{R}}(w_0) \leq 1$, and alternatively invoking in Theorem 3.4. QED.

If Z_c is empty, the proof is complete by plugging $\eta_j = 1/\sqrt{j+1}$ into Lemma 3.19. The rest of this subsection assumes Z_c is nonempty and establishes convergence to $\bar{u} \in S^\perp$.

First we prove the following key lemma, using the Fenchel-Young inequality.

Lemma 3.20. For any $w \in \mathbb{R}^d$,

$$\frac{\langle \bar{u}, w \rangle}{\|w\|_2} \geq \frac{-\ln\left(\widehat{\mathcal{R}}(w) - \bar{\mathcal{R}}\right)}{\gamma \|w\|_2} - \frac{\ln(n) + \left\|\Pi_S(w)\right\|_2}{\gamma \|w\|_2}.$$

Note the additional term $\left\|\Pi_S(w)\right\|_2$; by Lemma 3.19, this term is bounded.

Proof. First note that

$$Z_\perp w = Z_c \Pi_\perp w = Z_c w + Z_c(\Pi_\perp w - w) = Z_c w - Z_c \Pi_S w,$$

thus

$$\begin{aligned} \langle \bar{q}, Z_\perp w \rangle &= \langle \bar{q}, Z_c w \rangle - \left\langle \bar{q}, Z_c \Pi_S(w) \right\rangle \geq \langle \bar{q}, Z_c w \rangle - \|\bar{q}\|_1 \|Z_c \Pi_S(w)\|_\infty \\ &= \langle \bar{q}, Z_c w \rangle - \max_i (Z_c)_{i:} \Pi_S(w) \geq \langle \bar{q}, Z_c w \rangle - \left\|\Pi_S(w)\right\|_2, \end{aligned}$$

where $\bar{q}$ is the optimal dual solution given by Theorem 3.3. Moreover, note that given $\xi \in \mathbb{R}^k$, we have $\ln L(\xi) = \ln \sum_{i=1}^k \exp(-\xi_i)$, so it follows from the Fenchel-Young inequality that

$$\begin{aligned}
\frac{\langle \bar{u}, w \rangle}{\|w\|_2} = \frac{\left\langle Z_\perp^\top \bar{q}, w \right\rangle}{\gamma \|w\|_2} = \frac{\langle \bar{q}, Z_\perp w \rangle}{\gamma \|w\|_2} &\geq -\frac{\langle \bar{q}, -Z_c w \rangle}{\gamma \|w\|_2} - \frac{\left\|\Pi_S(w)\right\|_2}{\gamma \|w\|_2} \\
&\geq -\frac{\ln L_c(w) + g^*(\bar{q})}{\gamma \|w\|_2} - \frac{\left\|\Pi_S(w)\right\|_2}{\gamma \|w\|_2} \\
&\geq -\frac{\ln \widehat{\mathcal{R}}_c(w) + \ln(n)}{\gamma \|w\|_2} - \frac{\left\|\Pi_S(w)\right\|_2}{\gamma \|w\|_2},
\end{aligned}$$

where $g^*(\bar{q}) \leq 0$ denotes the negative entropy of $\bar{q}$. Also note that $\widehat{\mathcal{R}}_c(w) \leq \widehat{\mathcal{R}}(w) - \bar{\mathcal{R}}$, we have

$$\frac{\langle \bar{u}, w \rangle}{\|w\|_2} \geq -\frac{\ln\left(\widehat{\mathcal{R}}(w) - \bar{\mathcal{R}}\right) + \ln(n)}{\gamma \|w\|_2} - \frac{\left\|\Pi_S(w)\right\|_2}{\gamma \|w\|_2}.$$

QED.

To apply Lemma 3.20, we also need the following bound on $\widehat{\mathcal{R}}(w_t) - \bar{\mathcal{R}}$.

Lemma 3.21. Suppose $\hat{\eta}_j \leq 1$ (meaning $\eta_j \leq 1/\widehat{\mathcal{R}}(w_j)$). Also suppose that j is large enough such that $\widehat{\mathcal{R}}(w_j) - \bar{\mathcal{R}} \leq \lambda(1-r)/2$ for some $r \in (0,1)$, where λ is the strong convexity modulus of $\widehat{\mathcal{R}}_S$ over the 1-sublevel set. Then

$$\widehat{\mathcal{R}}(w_{j+1}) - \bar{\mathcal{R}} \leq \left(\widehat{\mathcal{R}}(w_j) - \bar{\mathcal{R}}\right) \exp\left(-r\gamma\gamma_j\hat{\eta}_j\left(1 - \hat{\eta}_j/2\right)\right).$$

Proof of Lemma 3.21. Invoking Lemmas 3.17 and 3.18 and proceeding as in Lemma 3.18,

$$\begin{aligned}
\widehat{\mathcal{R}}(w_{j+1}) - \bar{\mathcal{R}} &\leq \widehat{\mathcal{R}}(w_j) - \bar{\mathcal{R}} - \eta_j \left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2^2 + \frac{\eta_j^2 \widehat{\mathcal{R}}(w_j)}{2} \left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2^2 \\
&\leq \left(\widehat{\mathcal{R}}(w_j) - \bar{\mathcal{R}}\right) \left(1 - \frac{\eta_j \left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2^2}{\widehat{\mathcal{R}}(w_j) - \bar{\mathcal{R}}} \left(1 - \eta_j \widehat{\mathcal{R}}(w_j)/2\right)\right) \\
&= \left(\widehat{\mathcal{R}}(w_j) - \bar{\mathcal{R}}\right) \left(1 - \frac{\left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2}{\widehat{\mathcal{R}}(w_j) - \bar{\mathcal{R}}} \cdot \frac{\eta_j \widehat{\mathcal{R}}(w_j) \left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2}{\widehat{\mathcal{R}}(w_j)} \left(1 - \eta_j \widehat{\mathcal{R}}(w_j)/2\right)\right) \\
&\leq \left(\widehat{\mathcal{R}}(w_j) - \bar{\mathcal{R}}\right) \left(1 - \frac{\left\|\nabla \widehat{\mathcal{R}}(w_j)\right\|_2}{\widehat{\mathcal{R}}(w_j) - \bar{\mathcal{R}}} \cdot \hat{\eta}_j \gamma_j \left(1 - \hat{\eta}_j/2\right)\right). \qquad (3.21)
\end{aligned}$$

Next it will be shown, by analyzing two cases, that

$$\frac{\left\|\nabla\widehat{\mathcal{R}}(w_j)\right\|_2}{\widehat{\mathcal{R}}(w_j)-\bar{\mathcal{R}}} \geq r\gamma. \tag{3.22}$$

In the following, for notational simplicity let w denote w_j.

- Suppose $\widehat{\mathcal{R}}_c(w) < r\left(\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}\right)$. Consequently,

$$\widehat{\mathcal{R}}_S(w)-\bar{\mathcal{R}} > (1-r)\left(\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}\right).$$

  Then, since $4\left(\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}\right) \leq 2\lambda(1-r)$,

$$\begin{aligned}
\frac{\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2}{\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}} &\geq \frac{-\left\|\nabla\widehat{\mathcal{R}}_c(w)\right\|_2+\left\|\nabla\widehat{\mathcal{R}}_S(w)\right\|_2}{\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}} \\
&\geq \frac{-\widehat{\mathcal{R}}_c(w)+\sqrt{2\lambda(\widehat{\mathcal{R}}_S(w)-\bar{\mathcal{R}})}}{\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}} \\
&> \frac{-r(\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}})+\sqrt{2\lambda(1-r)(\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}})}}{\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}} \\
&\geq \frac{(2-r)(\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}})}{\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}} \\
&\geq 1 \geq r\gamma.
\end{aligned}$$

- Otherwise, suppose $\widehat{\mathcal{R}}_c(w) \geq r\left(\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}\right)$. Using an expression inspired by a general analysis of AdaBoost [71, Lemma 16 of journal version],

$$\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2 \geq \langle\bar{u},\nabla\widehat{\mathcal{R}}(w)\rangle = \langle Z\bar{u},\nabla L(Zw)/n\rangle = \langle Z_c\bar{u},\nabla L(Z_cw)/n\rangle \geq \gamma\widehat{\mathcal{R}}_c(w),$$

  Thus

$$\frac{\left\|\nabla\widehat{\mathcal{R}}(w)\right\|_2}{\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}} \geq \frac{\gamma\widehat{\mathcal{R}}_c(w)}{\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}} \geq \frac{\gamma\left(r\left(\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}\right)\right)}{\widehat{\mathcal{R}}(w)-\bar{\mathcal{R}}} = r\gamma.$$

Combining eq. (3.21) with eq. (3.22),

$$\widehat{\mathcal{R}}(w_{j+1})-\bar{\mathcal{R}} \leq \left(\widehat{\mathcal{R}}(w_j)-\bar{\mathcal{R}}\right)\left(1-r\gamma\gamma_j\hat{\eta}_j\left(1-\hat{\eta}_j/2\right)\right).$$

QED.

A key property of the upper bound in Lemma 3.21 is that it has replaced γ_j^2 in Lemma 3.16 with $\gamma_j\gamma$. Plugging this bound into the Fenchel-Young scheme in Lemma 3.20 will now fortuitously cancel γ, which leads to the following promising bound.

Lemma 3.22. Select t_0 so that $\widehat{\mathcal{R}}(w_{t_0}) - \bar{\mathcal{R}} \leq \min\{1/n, \lambda(1-r)/2\}$ for some $r \in (0,1)$. Then for any $t \geq t_0$ and any w,

$$\frac{\langle \bar{u}, w_t \rangle}{\|w_t\|_2} \geq \frac{r\sum_{j=t_0}^{t-1} \hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j}{\|w_t\|_2} - \frac{\left\|\Pi_S(w_t)\right\|_2}{\gamma\|w_t\|_2}.$$

Proof. By Lemma 3.16, since $\eta_j \leq 1$, the loss decreases at each step, and thus for any $t \geq t_0$, $\widehat{\mathcal{R}}(w_t) \leq \lambda(1-r)/2$. Combining Lemma 3.20 and Lemma 3.21,

$$\begin{aligned}
\frac{\langle \bar{u}, w_t \rangle}{\|w_t\|_2} &\geq \frac{-\ln\left(\widehat{\mathcal{R}}(w_t) - \bar{\mathcal{R}}\right)}{\gamma\|w_t\|_2} - \frac{\ln(n) + \left\|\Pi_S(w_t)\right\|_2}{\gamma\|w_t\|_2}. \\
&\geq \frac{r\gamma\sum_{j=t_0}^{t-1} \hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j - \ln\left(\widehat{\mathcal{R}}(w_{t_0}) - \bar{\mathcal{R}}\right)}{\gamma\|w_t\|_2} - \frac{\ln(n) + \left\|\Pi_S(w_t)\right\|_2}{\gamma\|w_t\|_2} \\
&\geq \frac{r\sum_{j=t_0}^{t-1} \hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j}{\|w_t\|_2} - \frac{\ln(1/n)}{\gamma\|w_t\|_2} - \frac{\ln(n) + \left\|\Pi_S(w)\right\|_2}{\gamma\|w_t\|_2} \\
&\geq \frac{r\sum_{j=t_0}^{t-1} \hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j}{\|w_t\|_2} - \frac{\left\|\Pi_S(w)\right\|_2}{\gamma\|w_t\|_2}.
\end{aligned}$$

QED.

Now we are ready to prove Theorem 3.5.

Proof of Theorem 3.5. The guarantee on $\bar{v}_t$ and the $Z_c = \emptyset$ case have been discussed in Lemma 3.19, therefore assume $Z_c \neq \emptyset$. The proof will proceed via invocation of the Fenchel-Young scheme in Lemma 3.22, applied to w_t.

It is necessary to first control the warm start parameter t_0. Fix an arbitrary $\epsilon \in (0,1)$, set $r := 1 - \epsilon/2$, and let t_0 be large enough such that

$$\widehat{\mathcal{R}}(w_{t_0}) - \bar{\mathcal{R}} \leq \frac{1}{n}, \qquad \widehat{\mathcal{R}}(w_{t_0}) - \bar{\mathcal{R}} \leq \frac{\lambda(1-r)}{2} = \frac{\lambda\epsilon}{4}, \qquad 1 - \frac{\hat{\eta}_{t_0}}{2} \geq 1 - \frac{\eta_{t_0}}{2} \geq 1 - \frac{\epsilon}{2}. \tag{3.23}$$

By Theorem 3.4 and the choice of step sizes, it is enough to require

$$\frac{\exp(\|\bar{v}\|_2)}{t_0} \leq \min\left\{\frac{1}{2n}, \frac{\lambda\epsilon}{8}\right\}, \qquad \frac{\|\bar{v}\|_2^2 + \ln(t_0)^2/\gamma^2}{2\sqrt{t_0}} \leq \min\left\{\frac{1}{2n}, \frac{\lambda\epsilon}{8}\right\}, \qquad \frac{1}{2\sqrt{t_0+1}} \leq \frac{\epsilon}{2}.$$

Therefore, choosing $t_0 = O\left(\max\left\{n^2, 1/\epsilon^2\right\}\right)$ suffices.

Invoking Lemma 3.22 with the above choice for w_t,

$$\begin{aligned}
\frac{1}{2}\left\| \frac{w_t}{\|w_t\|_2} - \bar{u} \right\|_2^2 &= 1 - \frac{\langle \bar{u}, w \rangle}{\|w_t\|_2} \\
&\leq 1 - \frac{r \sum_{j=t_0}^{t-1} \hat{\eta}_j (1 - \hat{\eta}_j/2)\gamma_j}{\|w_t\|_2} + \frac{\left\|\Pi_S(w)\right\|_2}{\gamma \|w_t\|_2} \\
&\leq 1 - \frac{(1-\epsilon/2) \sum_{j=t_0}^{t-1} \hat{\eta}_j (1 - \epsilon/2)\gamma_j}{\|w_t\|_2} + \frac{\left\|\Pi_S(w)\right\|_2}{\gamma \|w_t\|_2} \\
&\leq 1 - \frac{(1-\epsilon) \sum_{j=t_0}^{t-1} \hat{\eta}_j \gamma_j}{\|w_t\|_2} + \frac{\left\|\Pi_S(w)\right\|_2}{\gamma \|w_t\|_2} \\
&= 1 - \frac{(1-\epsilon)\left(\|w_{t_0}\|_2 + \sum_{j=t_0}^{t-1} \hat{\eta}_j \gamma_j)\right)}{\|w_t\|_2} + (1-\epsilon)\frac{\|w_{t_0}\|_2}{\|w_t\|_2} + \frac{\left\|\Pi_S(w)\right\|_2}{\gamma \|w_t\|_2} \\
&\leq \epsilon + \frac{\|w_{t_0}\|_2}{\|w_t\|_2} + \frac{\left\|\Pi_S(w)\right\|_2}{\gamma \|w_t\|_2}.
\end{aligned} \tag{3.24}$$

Since ϵ is arbitrary, it follows that $w_t/\|w_t\|_2 \to \bar{u}$. QED.

Chapter 4: Wide two-layer ReLU networks

Despite the extensive empirical success of deep networks, their optimization and generalization properties are still not fully understood. Recently, the neural tangent kernel (NTK) has provided the following insight into the problem. In the infinite-width limit, the NTK converges to a limiting kernel which stays constant during training; on the other hand, when the width is large enough, the function learned by gradient descent follows the NTK [28]. This motivates the study of overparameterized networks trained by gradient descent, using properties of the NTK. In fact, parameters related to the NTK, such as the minimum eigenvalue of the limiting kernel, appear to affect optimization and generalization [72].

However, in addition to such NTK-dependent parameters, prior work also requires the width to depend polynomially on n, $1/\delta$ or $1/\epsilon$, where n denotes the size of the training set, δ denotes the failure probability, and ϵ denotes the target error. These large widths far exceed what is used empirically, constituting a significant gap between theory and practice.

In this chapter, we narrow this gap by showing that a two-layer ReLU network with $\Omega(\ln(n/\delta) + \ln(1/\epsilon)^2)$ hidden units trained by gradient descent achieves classification error ϵ *on test data*, meaning both optimization and generalization occur. Unlike prior work, the width is fully polylogarithmic in n, $1/\delta$, and $1/\epsilon$; the width will additionally depend on the *separation margin* of the limiting kernel, a quantity which is guaranteed positive (assuming no inputs are parallel), can distinguish between true labels and random labels, and can give a tight sample-complexity analysis in the infinite-width setting. The chapter organization together with some details are described below.

Section 4.1 studies gradient descent on the training set. Using the ℓ_1 geometry inherent in classification tasks, we prove that with any width at least polylogarithmic and any constant step size no larger than 1, gradient descent achieves training error ϵ in $\widetilde{\Theta}(1/\epsilon)$ iterations (cf. Theorem 4.1). As is common in the NTK literature [73], we also show the parameters hardly change, which will be essential to our generalization analysis.

Section 4.2 gives a test error bound. Concretely, using the preceding gradient descent analysis, and standard Rademacher tools and exploiting how little the weights moved, we show that with $\widetilde{\Omega}(1/\epsilon^2)$ samples and $\widetilde{\Theta}(1/\epsilon)$ iterations, gradient descent finds a solution with ϵ test error (cf. Theorem 4.2 and Corollary 4.1). (As discussed in Remark 4.1, $\widetilde{\Omega}(1/\epsilon)$ samples also suffice via a smoothness-based generalization bound, at the expense of large constant factors.)

Section 4.3 considers stochastic gradient descent (SGD) with access to a standard stochas-

tic online oracle. We prove that with width at least polylogarithmic and $\widetilde{\Theta}(1/\epsilon)$ samples, SGD achieves an arbitrarily small test error (cf. Theorem 4.3).

Section 4.4 discusses the separation margin, which is in general a positive number, but reflects the difficulty of the classification problem in the infinite-width limit. While this margin can degrade all the way down to $O(1/\sqrt{n})$ for random labels (cf. Proposition 4.5), it can be much larger when there is a strong relationship between features and labels: for example, on the *noisy 2-XOR* data introduced in [74], we show that the margin is $\Omega(1/d)$, which further implies our SGD sample complexity is tight in the infinite-width case (cf. Section 4.4.2).

The contents of this chapter is based on [30].

Related work. There has been a large literature studying gradient descent on overparameterized networks via the NTK. The most closely related work is [75], which showed that a two-layer network trained by gradient descent with the logistic loss can achieve a small test error, under the same assumption that the NTK with respect to the first layer can separate the data distribution. However, they analyzed smooth activations, while we handle the ReLU. They required $\Omega(1/\epsilon^2)$ hidden units, $\widetilde{\Omega}(1/\epsilon^4)$ data samples, and $O(1/\epsilon^2)$ steps, while our result only needs polylogarithmic hidden units, $\widetilde{\Omega}(1/\epsilon^2)$ data samples, and $\widetilde{O}(1/\epsilon)$ steps.

Additionally on shallow networks, [76] proved that on an overparameterized two-layer network, gradient descent can globally minimize the empirical risk with the squared loss. Their result requires $\Omega(n^6/\delta^3)$ hidden units. [77, 78] further reduced the required overparameterization, but there is still a $\mathrm{poly}(n)$ dependency. Using the same amount of overparameterization as [76], [72] further showed that the two-layer network learned by gradient descent can achieve a small test error, assuming that on the data distribution the smallest eigenvalue of the limiting kernel is at least some positive constant. They also gave a fine-grained characterization of the predictions made by gradient descent iterates; such a characterization makes use of a special property of the squared loss and cannot be applied to the logistic regression setting. [79] showed that stochastic gradient descent (SGD) with the cross entropy loss can learn a two-layer network with small test error, using $\mathrm{poly}(\ell, 1/\epsilon)$ hidden units, where ℓ is at least the covering number of the support of the feature distribution using balls whose radii are no larger than the smallest distance between two data points with different labels. [80] considered SGD on a two-layer network, and a variant of SGD on a three-layer network. The three-layer analysis further exhibits some properties not captured by the NTK. They assume a ground truth network with infinite-order smooth activations, and they require the

width to depend polynomially on $1/\epsilon$ and some constants related to the smoothness of the activations of the ground truth network.

On deep networks, a variety of works have established low training error [81, 82, 83, 84]. [85] showed that SGD can minimize the regression loss for recurrent neural networks, and [86] further proved a low generalization error. [87] showed that using the same number of training examples, a three-layer ResNet can learn a function class with a much lower test error than any kernel method. [36] assumed that the NTK with respect to the second layer of a two-layer network can separate the data distribution, and proved that gradient descent on a deep network can achieve ϵ test error with $\Omega(1/\epsilon^4)$ samples and $\Omega(1/\epsilon^{14})$ hidden units. [88] considered SGD with an online oracle and give a general result. Under the same assumption as in [36], their result requires $\Omega(1/\epsilon^{14})$ hidden units and sample complexity $\widetilde{O}(1/\epsilon^2)$. By contrast, with the same online oracle, our result only needs polylogarithmic hidden units and sample complexity $\widetilde{O}(1/\epsilon)$. [29] extended our analysis to deep networks.

Notation. As usual, the dataset is denoted by $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. For simplicity, we assume that $\|x_i\|_2 = 1$ for any $1 \le i \le n$, which is standard in the NTK literature.

The two-layer network has weight matrices $W \in \mathbb{R}^{m\times d}$ and $a \in \mathbb{R}^m$. We use the following parameterization, which is also used in [72, 76]:

$$f(x; W, a) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma\left(\langle w_s, x\rangle\right),$$

with initialization

$$w_{s,0} \sim \mathcal{N}(0, I_d), \quad \text{and} \quad a_s \sim \text{unif}\left(\{-1, +1\}\right).$$

Note that in this paper, $w_{s,t}$ denotes the s-th row of W at step t. We fix a and only train W, as in [72, 75, 76, 79]. We consider the ReLU activation $\sigma(z) := \max\{0, z\}$, though our analysis can be extended easily to Lipschitz continuous, positively homogeneous activations such as leaky ReLU.

For simplicity, in this chapter we let ℓ denote the logistic loss. For any $1 \le i \le n$ and any W, let $f_i(W) := f(x_i; W, a)$. The empirical risk and its gradient are given by

$$\widehat{\mathcal{R}}(W) := \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i f_i(W)\right), \quad \text{and} \quad \nabla\widehat{\mathcal{R}}(W) = \frac{1}{n} \sum_{i=1}^{n} \ell'\left(y_i f_i(W)\right) y_i \nabla f_i(W).$$

For any $t \geq 0$, the gradient descent step is given by $W_{t+1} := W_t - \eta_t \nabla \widehat{\mathcal{R}}(W_t)$. Also define

$$f_i^{(t)}(W) := \left\langle \nabla f_i(W_t), W \right\rangle, \quad \text{and} \quad \widehat{\mathcal{R}}^{(t)}(W) := \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i f_i^{(t)}(W)\right).$$

Note that $f_i^{(t)}(W_t) = f_i(W_t)$. This property generally holds due to homogeneity: for any W and any $1 \leq s \leq m$,

$$\frac{\partial f_i}{\partial w_s} = \frac{1}{\sqrt{m}} a_s \mathbb{1}\left[\langle w_s, x_i \rangle > 0\right] x_i, \quad \text{and} \quad \left\langle \frac{\partial f_i}{\partial w_s}, w_s \right\rangle = \frac{1}{\sqrt{m}} a_s \sigma\left(\langle w_s, x_i \rangle\right),$$

and thus $\left\langle \nabla f_i(W), W \right\rangle = f_i(W)$.

4.1 EMPIRICAL RISK MINIMIZATION

In this section, we consider a fixed training set and empirical risk minimization. We first state our assumption on the separability of the NTK, and then give our main result and a proof sketch.

The key idea of the NTK is to do the first-order Taylor approximation:

$$f(x; W, a) \approx f(x; W_0, a) + \left\langle \nabla_W f(x; W_0, a), W - W_0 \right\rangle.$$

In other words, we want to do learning using the features given by $\nabla f_i(W_0) \in \mathbb{R}^{m \times d}$. A natural assumption is that there exists $\overline{U} \in \mathbb{R}^{m \times d}$ which can separate $\left\{\left(\nabla f_i(W_0), y_i\right)\right\}_{i=1}^n$ with a positive margin:

$$\min_{1 \leq i \leq n} \left(y_i \left\langle \overline{U}, \nabla f_i(W_0) \right\rangle \right) = \min_{1 \leq i \leq n} \left(y_i \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \langle \bar{u}_s, x_i \rangle \mathbb{1}\left[\langle w_{s,0}, x_i \rangle > 0\right] \right) > 0. \tag{4.1}$$

The infinite-width limit of eq. (4.1) is formalized as Assumption 4.1, with an additional bound on the $(2, \infty)$ norm of the separator. A concrete construction of $\overline{U}$ using Assumption 4.1 is given in eq. (4.2).

Let $\mu_{\mathcal{N}}$ denote the Gaussian measure on $\mathbb{R}^d$, given by the Gaussian density with respect to the Lebesgue measure on $\mathbb{R}^d$. We consider the following Hilbert space

$$\mathcal{H} := \left\{ w : \mathbb{R}^d \to \mathbb{R}^d \;\middle|\; \int \|w(z)\|_2^2 \, \mathrm{d}\mu_{\mathcal{N}}(z) < \infty \right\}.$$

For any $x \in \mathbb{R}^d$, define $\phi_x \in \mathcal{H}$ by $\phi_x(z) := x \mathbb{1}\left[\langle z, x \rangle > 0\right]$, and particularly define $\phi_i := \phi_{x_i}$

for the training input x_i.

Assumption 4.1. There exists $\bar{v} \in \mathcal{H}$ and $\gamma > 0$, such that $\left\|\bar{v}(z)\right\|_2 \leq 1$ for any $z \in \mathbb{R}^d$, and for any $1 \leq i \leq n$,

$$y_i \left\langle \bar{v}, \phi_i \right\rangle_{\mathcal{H}} := y_i \int \left\langle \bar{v}(z), \phi_i(z) \right\rangle \mathrm{d}\mu_{\mathcal{N}}(z) \geq \gamma.$$

As discussed in Section 4.4, the space $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) induced by the infinite-width NTK with respect to W, and ϕ_x maps x into $\mathcal{H}$. Assumption 4.1 supposes that the induced training set $\{(\phi_i, y_i)\}_{i=1}^n$ can be separated by some $\bar{v} \in \mathcal{H}$, with an additional bound on $\left\|\bar{v}(z)\right\|_2$ which is crucial in our analysis. It is also possible to give a dual characterization of the separation margin (cf. eq. (4.18)), which also allows us to show that Assumption 4.1 always holds when there are no parallel inputs (cf. Proposition 4.4). However, it is often more convenient to construct $\bar{v}$ directly; see Section 4.4 for some examples.

With Assumption 4.1, we state our main empirical risk result.

Theorem 4.1. Under Assumption 4.1, given any risk target $\epsilon \in (0, 1)$ and any $\delta \in (0, 1/3)$, let

$$\lambda := \frac{\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)}{\gamma/4}, \quad \text{and} \quad M := \frac{4096\lambda^2}{\gamma^6}.$$

Then for any $m \geq M$ and any constant step size $\eta \leq 1$, with probability $1 - 3\delta$ over the random initialization,

$$\frac{1}{T}\sum_{t<T} \widehat{\mathcal{R}}(W_t) \leq \epsilon, \quad \text{where} \quad T := \left\lceil \frac{2\lambda^2}{\eta\epsilon} \right\rceil.$$

Moreover for any $0 \leq t < T$ and any $1 \leq s \leq m$,

$$\left\|w_{s,t} - w_{s,0}\right\|_2 \leq \frac{4\lambda}{\gamma\sqrt{m}}.$$

While the number of hidden units required by prior work all have a polynomial dependency on n, $1/\delta$ or $1/\epsilon$, Theorem 4.1 only requires $m = \Omega\left(\ln(n/\delta) + \ln(1/\epsilon)^2\right)$. The required width has a polynomial dependency on $1/\gamma$, which is an adaptive quantity: while $1/\gamma$ can be $\mathrm{poly}(n)$ for random labels (cf. Proposition 4.5), it can be $\mathrm{polylog}(n)$ when there is a strong feature-label relationship, for example on the noisy 2-XOR data introduced in [74] (cf. Proposition 4.6). Moreover, we show in Proposition 4.7 that if we want $\left\{\left(\nabla f_i(W_0), y_i\right)\right\}_{i=1}^n$

to be separable, which is the starting point of an NTK-style analysis, the width has to depend polynomially on $1/\gamma$.

In the rest of Section 4.1, we prove Theorem 4.1.

4.1.1 Properties at initialization

In this subsection, we give some nice properties of random initialization.

Given an initialization (W_0, a), for any $1 \le s \le m$, define

$$\bar{u}_s := \frac{1}{\sqrt{m}} a_s \bar{v}(w_{s,0}), \tag{4.2}$$

where $\bar{v}$ is given by Assumption 4.1. Collect $\bar{u}_s$ into a matrix $\overline{U} \in \mathbb{R}^{m \times d}$. It holds that $\|\bar{u}_s\|_2 \le 1/\sqrt{m}$, and $\left\|\overline{U}\right\|_F \le 1$.

Lemma 4.1 ensures that, with high probability, $\overline{U}$ constructed above has a positive margin at initialization.

Lemma 4.1. Under Assumption 4.1, given any $\delta \in (0,1)$ and any $\epsilon_1 \in (0, \gamma)$, if $m \ge \left(2\ln(n/\delta)\right)/\epsilon_1^2$, then with probability $1-\delta$, it holds simultaneously for all $1 \le i \le n$ that

$$y_i f_i^{(0)}\left(\overline{U}\right) = y_i \left\langle \nabla f_i(W_0), \overline{U} \right\rangle \ge \gamma - \sqrt{\frac{2\ln(n/\delta)}{m}} \ge \gamma - \epsilon_1.$$

Proof. By Assumption 4.1, given any $1 \le i \le n$,

$$\mu := \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} \left[y_i \left\langle \bar{v}(w), x_i \right\rangle \mathbb{1}\left[\langle w, x_i \rangle > 0\right] \right] \ge \gamma.$$

On the other hand,

$$y_i f_i^{(0)}\left(\overline{U}\right) = \frac{1}{m} \sum_{s=1}^{m} y_i \left\langle \bar{v}(w_{s,0}), x_i \right\rangle \mathbb{1}\left[\left\langle w_{s,0}, x_i \right\rangle > 0\right]$$

is the empirical mean of i.i.d. r.v.'s supported on $[-1, +1]$ with mean μ. Therefore by Hoeffding's inequality, with probability $1 - \delta/n$,

$$y_i f_i^{(0)}\left(\overline{U}\right) - \gamma \ge y_i f_i^{(0)}\left(\overline{U}\right) - \mu \ge -\sqrt{\frac{2\ln(n/\delta)}{m}}.$$

Applying a union bound finishes the proof. QED.

For any W, any $\epsilon_2 > 0$, and any $1 \le i \le n$, define

$$\alpha_i(W, \epsilon_2) = \frac{1}{m} \sum_{s=1}^{m} \mathbb{1}\left[\left|\langle w_s, x_i \rangle\right| \le \epsilon_2\right].$$

Lemma 4.2 controls $\alpha_i(W_0, \epsilon_2)$. It will help us show that $\overline{U}$ has a good margin during the training process.

Lemma 4.2. Under the condition of Lemma 4.1, for any $\epsilon_2 > 0$, with probability $1 - \delta$, it holds simultaneously for all $1 \le i \le n$ that

$$\alpha_i\left(W_0, \epsilon_2\right) \le \sqrt{\frac{2}{\pi}}\epsilon_2 + \sqrt{\frac{\ln(n/\delta)}{2m}} \le \epsilon_2 + \frac{\epsilon_1}{2}.$$

Proof. Given any fixed ϵ_2 and $1 \le i \le n$,

$$\mathbb{E}\left[\alpha_i(W_0, \epsilon_2)\right] = \mathbb{P}\left(\left|\langle w, x_i \rangle\right| \le \epsilon_2\right) \le \frac{2\epsilon_2}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}\epsilon_2,$$

because $\langle w, x_i \rangle$ is a standard Gaussian r.v. and the density of standard Gaussian has maximum $1/\sqrt{2\pi}$. Since $\alpha_i(W_0, \epsilon_2)$ is the empirical mean of Bernoulli r.v.'s, by Hoeffding's inequality, with probability $1 - \delta/n$,

$$\alpha_i(W_0, \epsilon_2) \le \mathbb{E}\left[\alpha_i(W_0, \epsilon_2)\right] + \sqrt{\frac{\ln(n/\delta)}{2m}} \le \sqrt{\frac{2}{\pi}}\epsilon_2 + \sqrt{\frac{\ln(n/\delta)}{2m}}.$$

Applying a union bound finishes the proof. QED.

Finally, Lemma 4.3 controls the output of the network at initialization.

Lemma 4.3. Given any $\delta \in (0, 1)$, if $m \ge 25 \ln(2n/\delta)$, then with probability $1 - \delta$, it holds simultaneously for all $1 \le i \le n$ that

$$\left|f(x_i; W_0, a)\right| \le \sqrt{2 \ln\left(4n/\delta\right)}.$$

To prove Lemma 4.3, we need the following technical result.

Lemma 4.4. Consider the random vector $X = (X_1, \ldots, X_m)$, where $X_i = \sigma(Z_i)$ for some $\sigma : \mathbb{R} \to \mathbb{R}$ that is 1-Lipschitz, and Z_i are i.i.d. standard Gaussian r.v.'s. Then the r.v. $\|X\|_2$ is 1-sub-Gaussian, and thus with probability $1 - \delta$,

$$\|X\|_2 - \mathbb{E}\left[\|X\|_2\right] \le \sqrt{2 \ln(1/\delta)}.$$

Proof. Given $a \in \mathbb{R}^m$, define

$$f(a) = \sqrt{\sum_{i=1}^{m} \sigma(a_i)^2} = \left\|\sigma(a)\right\|_2,$$

where $\sigma(a)$ is obtained by applying σ coordinate-wisely to a. For any $a, b \in \mathbb{R}^m$, by the triangle inequality, we have

$$\left|f(a) - f(b)\right| = \left|\left\|\sigma(a)\right\|_2 - \left\|\sigma(b)\right\|_2\right| \leq \left\|\sigma(a) - \sigma(b)\right\|_2 = \sqrt{\sum_{i=1}^{m} \left(\sigma(a_i) - \sigma(b_i)\right)^2},$$

and by further using the 1-Lipschitz continuity of σ, we have

$$\left|f(a) - f(b)\right| \leq \sqrt{\sum_{i=1}^{m} \left(\sigma(a_i) - \sigma(b_i)\right)^2} \leq \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2} = \left\|a - b\right\|_2.$$

As a result, f is a 1-Lipschitz continuous function w.r.t. the ℓ_2 norm, indeed $f(X)$ is 1-sub-Gaussian and the bound follows by Gaussian concentration [89, Theorem 2.4]. QED.

Now we can prove Lemma 4.3

Proof of Lemma 4.3. Given $1 \leq i \leq n$, let $h_i = \sigma(W_0 x_i)/\sqrt{m}$. By Lemma 4.4, $\|h_i\|_2$ is sub-Gaussian with variance proxy $1/m$, and with probability at least $1 - \delta/(2n)$ over W_0,

$$\|h_i\|_2 - \mathbb{E}\left[\|h_i\|_2\right] \leq \sqrt{\frac{2\ln(2n/\delta)}{m}} \leq \sqrt{\frac{2\ln(2n/\delta)}{25\ln(2n/\delta)}} \leq 1 - \frac{\sqrt{2}}{2}.$$

On the other hand, by Jensen's inequality,

$$\mathbb{E}\left[\|h_i\|_2\right] \leq \sqrt{\mathbb{E}\left[\|h_i\|_2^2\right]} = \frac{\sqrt{2}}{2}.$$

As a result, with probability $1 - \delta/(2n)$, it holds that $\|h_i\|_2 \leq 1$. By a union bound, with probability $1 - \delta/2$ over W_0, for all $1 \leq i \leq n$, we have $\|h_i\|_2 \leq 1$.

For any W_0 such that the above event holds, and for any $1 \leq i \leq n$, the r.v. $\langle h_i, a\rangle$ is sub-Gaussian with variance proxy $\|h_i\|_2^2 \leq 1$. By Hoeffding's inequality, with probability $1 - \delta/(2n)$ over a,

$$\left|\langle h_i, a\rangle\right| = \left|f(x_i; W_0, a)\right| \leq \sqrt{2\ln\left(4n/\delta\right)}.$$

By a union bound, with probability $1-\delta/2$ over a, for all $1 \le i \le n$, we have $\left|f(x_i; W_0, a)\right| \le \sqrt{2\ln\left(4n/\delta\right)}$.

The probability that the above events all happen is at least $(1-\delta/2)(1-\delta/2) \ge 1-\delta$, over W_0 and a. QED.

4.1.2 Convergence analysis of gradient descent

We analyze gradient descent in this subsection. First, define

$$\widehat{\mathcal{Q}}(W) := \frac{1}{n}\sum_{i=1}^{n} -\ell'\left(y_i f_i(W)\right).$$

We have the following observations.

- For any W and any $1 \le s \le m$, $\left\|\partial f_i/\partial w_s\right\|_2 \le 1/\sqrt{m}$, and thus $\left\|\nabla f_i(W)\right\|_F \le 1$. Therefore by the triangle inequality, $\left\|\nabla\widehat{\mathcal{R}}(W)\right\|_F \le \widehat{\mathcal{Q}}(W)$.
- The logistic loss satisfies $0 \le -\ell' \le 1$, and thus $0 \le \widehat{\mathcal{Q}}(W) \le 1$.
- The logistic loss satisfies $-\ell' \le \ell$, and thus $\widehat{\mathcal{Q}}(W) \le \widehat{\mathcal{R}}(W)$.

The quantity $\widehat{\mathcal{Q}}$ first appeared in the perceptron analysis [10] for the ReLU loss, and has also been analyzed in prior work [14, 36, 75]. In this work, $\widehat{\mathcal{Q}}$ specifically helps us prove the following result, which plays an important role in obtaining a width which only depends on polylog$(1/\epsilon)$.

Lemma 4.5. For any $t \ge 0$ and any $\overline{W}$, if $\eta_t \le 1$, then

$$\eta_t\widehat{\mathcal{R}}(W_t) \le \left\|W_t - \overline{W}\right\|_F^2 - \left\|W_{t+1} - \overline{W}\right\|_F^2 + 2\eta_t\widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right).$$

Consequently, if we use a constant step size $\eta \le 1$ for $0 \le \tau < t$, then

$$\eta\left(\sum_{\tau<t}\widehat{\mathcal{R}}(W_\tau)\right) + \left\|W_t - \overline{W}\right\|_F^2 \le \left\|W_0 - \overline{W}\right\|_F^2 + 2\eta\left(\sum_{\tau<t}\widehat{\mathcal{R}}^{(\tau)}\left(\overline{W}\right)\right).$$

Lemma 4.5 is similar to [87, Fact D.4 and Claim D.5], where the squared loss is considered.

Proof of Lemma 4.5. We have

$$\left\|W_{t+1} - \overline{W}\right\|_F^2 = \left\|W_t - \overline{W}\right\|_F^2 - 2\eta_t\left\langle\nabla\widehat{\mathcal{R}}(W_t), W_t - \overline{W}\right\rangle + \eta_t^2\left\|\nabla\widehat{\mathcal{R}}(W_t)\right\|_F^2. \tag{4.3}$$

The first order term can be handled using the convexity of ℓ and homogeneity of ReLU:

$$\begin{aligned}\left\langle \nabla\widehat{\mathcal{R}}(W_t), W_t - \overline{W}\right\rangle &= \frac{1}{n}\sum_{i=1}^{n} \ell'\left(y_i f_i(W_t)\right) y_i \left\langle \nabla f_i(W_t), W_t - \overline{W}\right\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n} \ell'\left(y_i f_i(W_t)\right)\left(y_i f_i(W_t) - y_i f_i^{(t)}\left(\overline{W}\right)\right) \\ &\geq \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(y_i f_i(W_t)\right) - \ell\left(y_i f_i^{(t)}\left(\overline{W}\right)\right)\right) = \widehat{\mathcal{R}}(W_t) - \widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right).\end{aligned} \tag{4.4}$$

The second-order term of eq. (4.3) can be bounded as follows

$$\eta_t^2 \left\|\nabla\widehat{\mathcal{R}}(W_t)\right\|_F^2 \leq \eta_t^2 \widehat{\mathcal{Q}}(W_t)^2 \leq \eta_t \widehat{\mathcal{Q}}(W_t) \leq \eta_t \widehat{\mathcal{R}}(W_t), \tag{4.5}$$

because $\left\|\nabla\widehat{\mathcal{R}}(W_t)\right\|_F \leq \widehat{\mathcal{Q}}(W_t)$, and $\eta_t, \widehat{\mathcal{Q}}(W_t) \leq 1$, and $\widehat{\mathcal{Q}}(W_t) \leq \widehat{\mathcal{R}}(W_t)$. Combining eqs. (4.3) to (4.5) gives

$$\eta_t \widehat{\mathcal{R}}(W_t) \leq \left\|W_t - \overline{W}\right\|_F^2 - \left\|W_{t+1} - \overline{W}\right\|_F^2 + 2\eta_t \widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right).$$

Telescoping gives the other claim. QED.

Now we are ready to prove Theorem 4.1.

Proof of Theorem 4.1. The required width ensures that with probability $1-3\delta$, Lemmas 4.1 to 4.3 hold with $\epsilon_1 = \gamma^2/8$ and $\epsilon_2 = 4\lambda/(\gamma\sqrt{m})$.

Let t_1 denote the first step such that there exists $1 \leq s \leq m$ with $\left\|w_{s,t_1} - w_{s,0}\right\|_2 > 4\lambda/(\gamma\sqrt{m})$. Therefore for any $0 \leq t < t_1$ and any $1 \leq s \leq m$, it holds that $\left\|w_{s,t} - w_{s,0}\right\|_2 \leq 4\lambda/(\gamma\sqrt{m})$. In addition, we let $\overline{W} := W_0 + \lambda\overline{U}$.

We first prove that for any $0 \leq t < t_1$, it holds that $\widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right) \leq \epsilon/4$. Since the logistic satisfies $\ell(z) \leq \exp(-z)$, and it is enough to prove that for any $1 \leq i \leq n$,

$$y_i \left\langle \nabla f_i(W_t), \overline{W}\right\rangle \geq \ln\left(\frac{4}{\epsilon}\right).$$

We will split the left hand side into three terms and control them individually:

$$y_i \left\langle \nabla f_i(W_t), \overline{W}\right\rangle = y_i \left\langle \nabla f_i(W_0), W_0\right\rangle + y_i \left\langle \nabla f_i(W_t) - \nabla f_i(W_0), W_0\right\rangle + \lambda y_i \left\langle \nabla f_i(W_t), \overline{U}\right\rangle. \tag{4.6}$$

The first term of eq. (4.6) can be controlled using Lemma 4.3:

$$\left|y_i \left\langle \nabla f_i(W_0), W_0 \right\rangle\right| \leq \sqrt{2\ln(4n/\delta)}. \tag{4.7}$$

The second term of eq. (4.6) can be written as

$$y_i \left\langle \nabla f_i(W_t) - \nabla f_i(W_0), W_0 \right\rangle = \frac{y_i}{\sqrt{m}} \sum_{s=1}^{m} a_s \left(\mathbb{1}\left[\left\langle w_{s,t}, x_i \right\rangle > 0\right] - \mathbb{1}\left[\left\langle w_{s,0}, x_i \right\rangle > 0\right]\right) \left\langle w_{s,0}, x_i \right\rangle.$$

Let $S_c := \left\{ s \;\middle|\; \mathbb{1}\left[\left\langle w_{s,t}, x_i \right\rangle > 0\right] - \mathbb{1}\left[\left\langle w_{s,0}, x_i \right\rangle > 0\right] \neq 0 \right\}$. Note that $s \in S_c$ implies

$$\left|\left\langle w_{s,0}, x_i \right\rangle\right| \leq \left|\left\langle w_{s,t} - w_{s,0}, x_i \right\rangle\right| \leq \left\|w_{s,t} - w_{s,0}\right\|_2 \|x_i\|_2 = \left\|w_{s,t} - w_{s,0}\right\|_2 \leq 4\lambda/(\gamma\sqrt{m}) = \epsilon_2.$$

Therefore Lemma 4.2 ensures that

$$|S_c| \leq \left| \left\{ s \;\middle|\; |\left\langle w_{s,0}, x_i \right\rangle| \leq \epsilon_2 \right\} \right| \leq m\left(\frac{4\lambda}{\gamma\sqrt{m}} + \frac{\epsilon_1}{2}\right) = m\left(\frac{4\lambda}{\gamma\sqrt{m}} + \frac{\gamma^2}{16}\right).$$

and thus

$$\left|y_i \left\langle \nabla f_i(W_t) - \nabla f_i(W_0), W_0 \right\rangle\right| \leq \frac{1}{\sqrt{m}} \cdot |S_c| \cdot \frac{4\lambda}{\gamma\sqrt{m}} \leq \frac{16\lambda^2}{\gamma^2\sqrt{m}} + \frac{\lambda\gamma}{4} \leq \frac{\lambda\gamma}{2}, \tag{4.8}$$

where in the last step we use the condition that $m \geq 4096\lambda^2/\gamma^6$.

The third term of eq. (4.6) can be bounded as follows: by Lemma 4.1,

$$\begin{aligned} y_i \left\langle \nabla f_i(W_t), \overline{U} \right\rangle &= y_i \left\langle \nabla f_i(W_0), \overline{U} \right\rangle + y_i \left\langle \nabla f_i(W_t) - \nabla f_i(W_0), \overline{U} \right\rangle \\ &\geq \gamma - \epsilon_1 + y_i \left\langle \nabla f_i(W_t) - \nabla f_i(W_0), \overline{U} \right\rangle. \end{aligned}$$

In addition,

$$\begin{aligned} y_i \left\langle \nabla f_i(W_t) - \nabla f_i(W_0), \overline{U} \right\rangle &= \frac{y_i}{m} \sum_{i=1}^{m} \left(\mathbb{1}\left[\left\langle w_{s,t}, x_i \right\rangle > 0\right] - \mathbb{1}\left[\left\langle w_{s,0}, x_i \right\rangle > 0\right]\right) \left\langle \bar{v}(w_{s,0}), x_i \right\rangle \\ &\geq -\frac{1}{m} \cdot |S_c| \geq -\frac{4\lambda}{\gamma\sqrt{m}} - \frac{\epsilon_1}{2} \geq -\frac{\gamma^2}{16} - \frac{\epsilon_1}{2}, \end{aligned}$$

where we use $m \geq 4096\lambda^2/\gamma^6$. Therefore,

$$y_i \left\langle \nabla f_i(W_t), \overline{U} \right\rangle \geq \gamma - \epsilon_1 - \frac{\gamma^2}{16} - \frac{\epsilon_1}{2} = \gamma - \frac{\gamma^2}{4} \geq \frac{3\gamma}{4}. \tag{4.9}$$

Putting eqs. (4.7) to (4.9) into eq. (4.6), we have

$$y_i \left\langle \nabla f_i(W_t), \overline{W} \right\rangle \geq -\sqrt{2\ln\left(\frac{4n}{\delta}\right)} - \frac{\lambda\gamma}{2} + \frac{3\lambda\gamma}{4} = \frac{\lambda\gamma}{4} - \sqrt{2\ln\left(\frac{4n}{\delta}\right)} = \ln\left(\frac{4}{\epsilon}\right),$$

for the λ given in Theorem 4.1. Thus for any $0 \leq t < t_1$, it holds that $\widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right) \leq \epsilon/4$.

Let $T := \lceil 2\lambda^2/(\eta\epsilon) \rceil$, we claim that $t_1 \geq T$. To see this, note that Lemma 4.5 ensures

$$\left\| W_{t_1} - \overline{W} \right\|_F^2 \leq \left\| W_0 - \overline{W} \right\|_F^2 + 2\eta \left(\sum_{t<t_1} \widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right) \right) \leq \lambda^2 + \frac{\epsilon}{2}\eta t_1.$$

Suppose $t_1 < T$, then we have $t_1 \leq 2\lambda^2/(\eta\epsilon)$, and thus $\left\| W_{t_1} - \overline{W} \right\|_F^2 \leq 2\lambda^2$. As a result, using $\|\overline{U}\|_F \leq 1$ and the definition of $\overline{W}$,

$$\begin{aligned}\sqrt{2}\lambda \geq \left\| W_{t_1} - \overline{W} \right\|_F \geq \left\langle W_{t_1} - \overline{W}, \overline{U} \right\rangle &= \left\langle W_{t_1} - W_0, \overline{U} \right\rangle - \left\langle \overline{W} - W_0, \overline{U} \right\rangle \\ &\geq \left\langle W_{t_1} - W_0, \overline{U} \right\rangle - \lambda.\end{aligned}$$

Moreover, due to eq. (4.9),

$$\begin{aligned}\left\langle W_{t_1} - W_0, \overline{U} \right\rangle = -\eta \sum_{\tau<t_1} \left\langle \nabla \widehat{\mathcal{R}}(W_\tau), \overline{U} \right\rangle &= \eta \sum_{\tau<t_1} \frac{1}{n} \sum_{i=1}^n -\ell'\left(y_i f_i(W_\tau)\right) y_i \left\langle \nabla f_i(W_\tau), \overline{U} \right\rangle \\ &\geq \eta \sum_{\tau<t_1} \widehat{\mathcal{Q}}(W_\tau) \frac{3\gamma}{4}.\end{aligned}$$

which implies $\eta \sum_{\tau<t_1} \widehat{\mathcal{Q}}(W_\tau) \leq \frac{4(\sqrt{2}+1)\lambda}{3\gamma} \leq \frac{4\lambda}{\gamma}$. Furthermore, by the triangle inequality, for any $1 \leq s \leq m$,

$$\begin{aligned}\left\| w_{s,t} - w_{s,0} \right\|_2 &\leq \eta \sum_{\tau<t} \left\| \frac{1}{n} \sum_{i=1}^n \ell'\left(y_i f_i(W_\tau)\right) y_i \frac{\partial f_i}{\partial w_{s,\tau}} \right\|_2 \\ &\leq \eta \sum_{\tau<t} \frac{1}{n} \sum_{i=1}^n \left| \ell'\left(y_i f_i(W_\tau)\right) \right| \cdot \left\| \frac{\partial f_i}{\partial w_{s,\tau}} \right\|_2 \\ &\leq \eta \sum_{\tau<t} \widehat{\mathcal{Q}}(W_\tau) \frac{1}{\sqrt{m}} \leq \eta \sum_{\tau<t_1} \widehat{\mathcal{Q}}(W_\tau) \frac{1}{\sqrt{m}} \leq \frac{4\lambda}{\gamma\sqrt{m}}, \end{aligned} \tag{4.10}$$

which contradicts the definition of t_1. Therefore $t_1 \geq T$.

Now we are ready to prove the claims of Theorem 4.1. The bound on $\left\| w_{s,t} - w_{s,0} \right\|_2$ follow

by repeating the steps in eq. (4.10). The risk guarantee follows from Lemma 4.5:

$$\frac{1}{T}\sum_{t<T}\widehat{\mathcal{R}}(W_t) \le \frac{\left\|W_0 - \overline{W}\right\|_F^2}{\eta T} + \frac{2}{T}\sum_{t<T}\widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right) \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

QED.

4.2 GENERALIZATION

To get a generalization bound, we naturally extend Assumption 4.1 to the following assumption.

Assumption 4.2. There exists $\bar{v} \in \mathcal{H}$ and $\gamma > 0$, such that $\left\|\bar{v}(z)\right\|_2 \le 1$ for any $z \in \mathbb{R}^d$, and

$$y\int \left\langle \bar{v}(z), x\right\rangle \mathbb{1}\left[\langle z, x\rangle > 0\right] \mathrm{d}\mu_{\mathcal{N}}(z) \ge \gamma$$

for almost all (x, y) sampled from the data distribution $\mathcal{D}$.

The above assumption is also made in [75] for smooth activations. [36] make a similar separability assumption, but in the RKHS induced by the second layer a; by contrast, Assumption 4.2 is on separability in the RKHS induced by the first layer W.

Here is our test error bound with Assumption 4.2.

Theorem 4.2. Under Assumption 4.2, given any $\epsilon \in (0, 1)$ and any $\delta \in (0, 1/4)$, let λ and M be given as in Theorem 4.1:

$$\lambda := \frac{\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)}{\gamma/4}, \quad \text{and} \quad M := \frac{4096\lambda^2}{\gamma^6}.$$

Then for any $m \ge M$ and any constant step size $\eta \le 1$, with probability $1 - 4\delta$ over the random initialization and data sampling,

$$P_{(x,y)\sim\mathcal{D}}\left(yf(x; W_k, a) \le 0\right) \le 2\epsilon + \frac{16\left(\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)\right)}{\gamma^2\sqrt{n}} + 6\sqrt{\frac{\ln(2/\delta)}{2n}},$$

where k denotes the step with the minimum empirical risk before $\lceil 2\lambda^2/(\eta\epsilon)\rceil$.

Below is a direct corollary of Theorem 4.2.

Corollary 4.1. Under Assumption 4.2, given any $\epsilon, \delta \in (0,1)$, using a constant step size no larger than 1 and let

$$n = \widetilde{\Omega}\left(\frac{1}{\gamma^4\epsilon^2}\right), \quad \text{and} \quad m = \Omega\left(\frac{\ln(n/\delta) + \ln(1/\epsilon)^2}{\gamma^8}\right),$$

it holds with probability $1-\delta$ that $P_{(x,y)\sim\mathcal{D}}\left(yf(x;W_k,a) \le 0\right) \le \epsilon$, where k denotes the step with the minimum empirical risk in the first $\widetilde{\Theta}(1/(\gamma^2\epsilon))$ steps.

Remark 4.1. To get Theorem 4.2, we use a Lipschitz-based Rademacher complexity bound. One can also use a smoothness-based Rademacher complexity bound [90, Theorem 1] and get a sample complexity $\widetilde{O}(1/(\gamma^4\epsilon))$. However, the bound will become complicated and some large constant will be introduced. It is an interesting open question to give a clean analysis based on smoothness.

Below we prove Theorem 4.2; the key tool is a Rademacher complexity bound based on the sigmoid loss and $\left\|W_k^\top - W_0^\top\right\|_{2,\infty}$.

Given a sample $S = (z_1, \ldots, z_n)$ (where $z_i = (x_i, y_i)$) and a function class $\mathcal{H}$, the Rademacher complexity of $\mathcal{H}$ on S is defined as

$$\mathrm{Rad}\left(\mathcal{H} \circ S\right) := \frac{1}{n}\mathbb{E}_{\epsilon\sim\{-1,+1\}^n}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n \epsilon_i h(z_i)\right].$$

We will use the following general result.

Lemma 4.6. [91, Theorem 26.5] If $h(z) \in [a,b]$, then with probability $1-\delta$,

$$\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{z\sim\mathcal{D}}\left[h(z)\right] - \frac{1}{n}\sum_{i=1}^n h(z_i)\right) \le 2\mathrm{Rad}\left(\mathcal{H}\circ S\right) + 3(b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

We also need the following contraction lemma. Consider a feature sample $X = (x_1, \ldots, x_n)$ and a function class $\mathcal{F}$ on X. For each $1 \le i \le n$, let $g_i : \mathbb{R} \to \mathbb{R}$ denote a K-Lipschitz function. Let $g \circ \mathcal{F}$ denote the class of functions which map x_i to $g_i(f(x_i))$ for some $f \in \mathcal{F}$.

Lemma 4.7. [91, Lemma 26.9] $\mathrm{Rad}\left(g\circ\mathcal{F}\circ X\right) \le K\mathrm{Rad}\left(\mathcal{F}\circ X\right)$.

To prove Theorem 4.2, we need one more Rademacher complexity bound. Given a fixed initialization (W_0, a), consider the following classes:

$$\mathcal{W}_\rho := \left\{W \in \mathbb{R}^{m\times d} \;\middle|\; \left\|w_s - w_{s,0}\right\|_2 \le \rho \text{ for any } 1 \le s \le m\right\},$$

and

$$\mathcal{F}_\rho := \left\{ x \mapsto f(x; W, a) \;\middle|\; W \in \mathcal{W}_\rho \right\}.$$

Given a feature sample X, the following Lemma 4.8 controls the Rademacher complexity of $\mathcal{F}_\rho \circ X$. A similar version was given in [92, Theorem 43], and the proof is similar to the proof of [93, Theorem 18] which also pushes the supremum through and handles each hidden unit separately.

Lemma 4.8. $\operatorname{Rad}\left(\mathcal{F}_\rho \circ X\right) \leq \rho\sqrt{m/n}$.

Proof of Lemma 4.8. We have

$$\begin{aligned}
\mathbb{E}_\epsilon \left[\sup_{W \in \mathcal{W}_\rho} \sum_{i=1}^n \epsilon_i f(x_i; W, a) \right] &= \mathbb{E}_\epsilon \left[\sup_{W \in \mathcal{W}_\rho} \sum_{i=1}^n \epsilon_i \sum_{s=1}^m \frac{1}{\sqrt{m}} a_s \sigma\left(\langle w_s, x_i \rangle \right) \right] \\
&= \mathbb{E}_\epsilon \left[\frac{1}{\sqrt{m}} \sup_{W \in \mathcal{W}_\rho} \sum_{s=1}^m \sum_{i=1}^n \epsilon_i a_s \sigma\left(\langle w_s, x_i \rangle \right) \right] \\
&= \mathbb{E}_\epsilon \left[\frac{1}{\sqrt{m}} \sum_{s=1}^m \left(\sup_{\left\| w_s - w_{s,0} \right\|_2 \leq \rho} \sum_{i=1}^n \epsilon_i a_s \sigma\left(\langle w_s, x_i \rangle \right) \right) \right] \\
&= \frac{1}{\sqrt{m}} \sum_{i=1}^m \mathbb{E}_\epsilon \left[\sup_{\left\| w_s - w_{s,0} \right\|_2 \leq \rho} \sum_{i=1}^n \epsilon_i a_s \sigma\left(\langle w_s, x_i \rangle \right) \right].
\end{aligned}$$

Note that for any $1 \leq s \leq m$, the mapping $z \mapsto a_s \sigma(z)$ is 1-Lipschitz, and thus Lemma 4.7 gives

$$\begin{aligned}
\mathbb{E}_\epsilon \left[\sup_{W \in \mathcal{W}_\rho} \sum_{i=1}^n \epsilon_i f(x_i; W, a) \right] &\leq \frac{1}{\sqrt{m}} \sum_{i=1}^m \mathbb{E}_\epsilon \left[\sup_{\left\| w_s - w_{s,0} \right\|_2 \leq \rho} \sum_{i=1}^n \epsilon_i a_s \sigma\left(\langle w_s, x_i \rangle \right) \right] \\
&\leq \frac{1}{\sqrt{m}} \sum_{i=1}^m \mathbb{E}_\epsilon \left[\sup_{\left\| w_s - w_{s,0} \right\|_2 \leq \rho} \sum_{i=1}^n \epsilon_i \langle w_s, x_i \rangle \right].
\end{aligned}$$

Invoking the Rademacher complexity of linear classifiers [91, Lemma 26.10] then gives

$$\operatorname{Rad}\left(\mathcal{F}_\rho \circ X\right) = \frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{W \in \mathcal{W}_\rho} \sum_{i=1}^n \epsilon_i f(x_i; W, a) \right] \leq \frac{\rho\sqrt{m}}{\sqrt{n}}.$$

QED.

Now we are ready to prove the main generalization result Theorem 4.2.

Proof. Fix an initialization (W_0, a), and let $\mathcal{H} := \left\{(x,y) \mapsto -\ell'\left(yf(x)\right) \;\middle|\; f \in \mathcal{F}_\rho\right\}$. Since for any $h \in \mathcal{H}$ and any z, $h(z) \in [0,1]$, Lemma 4.6 ensures that with probability $1-\delta$ over the data sampling,

$$\sup_{h\in\mathcal{H}} \left(\mathbb{E}_{z\sim\mathcal{D}}\left[h(z)\right] - \frac{1}{n}\sum_{i=1}^{n} h(z_i) \right) = \sup_{W\in\mathcal{W}_\rho} \left(\mathcal{Q}(W) - \widehat{\mathcal{Q}}(W) \right) \le 2\mathrm{Rad}\left(\mathcal{H}\circ S\right) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Since for each $1 \le i \le n$, the mapping $z \mapsto -\ell'(y_i z)$ is $(1/4)$-Lipschitz, Lemma 4.7 further ensures that $\mathrm{Rad}\left(\mathcal{H}\circ S\right) \le \mathrm{Rad}\left(\mathcal{F}_\rho \circ X\right)/4$, and thus

$$\sup_{W\in\mathcal{W}_\rho} \left(\mathcal{Q}(W) - \widehat{\mathcal{Q}}(W) \right) \le \frac{\rho\sqrt{m}}{2\sqrt{n}} + 3\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{4.11}$$

On the other hand, Theorem 4.1 ensures that under the conditions of Theorem 4.2, for any fixed dataset, with probability $1-3\delta$ over the random initialization, we have

$$\widehat{\mathcal{Q}}(W_k) \le \widehat{\mathcal{R}}(W_k) \le \epsilon, \quad \text{and} \quad \left\|w_{s,k} - w_{s,0}\right\|_2 \le \frac{4\lambda}{\gamma\sqrt{m}}.$$

As a result, invoking eq. (4.11) with $\rho = 4\lambda/(\gamma\sqrt{m})$, with probability $1-4\delta$ over the random initialization and data sampling,

$$\mathcal{Q}(W_k) \le \widehat{\mathcal{Q}}(W_k) + \frac{2\lambda}{\gamma\sqrt{n}} + 3\sqrt{\frac{\ln(2/\delta)}{2n}} \le \epsilon + \frac{8\left(\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)\right)}{\gamma^2\sqrt{n}} + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Invoking $P_{(x,y)\sim\mathcal{D}}\left(yf(x;W,a) \le 0\right) \le 2\mathcal{Q}(W)$ finishes the proof. QED.

4.3 STOCHASTIC GRADIENT DESCENT

There are some different formulations of SGD. In this section, we consider SGD with an online oracle. We randomly sample W_0 and a, and fix a during training. At step i, a data example (x_i, y_i) is sampled from the data distribution. We still let $f_i(W) := f(x_i; W, a)$, and perform the following update

$$W_{i+1} := W_i - \eta_i \ell'\left(y_i f_i(W_i)\right) y_i \nabla f_i(W_i).$$

Note that here i starts from 0.

Still with Assumption 4.2, we show the following result.

Theorem 4.3. Under Assumption 4.2, given any $\epsilon, \delta \in (0,1)$, using a constant step size and $m = \Omega\left(\left(\ln(1/\delta) + \ln(1/\epsilon)^2\right)/\gamma^8\right)$, it holds with probability $1-\delta$ that

$$\frac{1}{n}\sum_{i=1}^{n} P_{(x,y)\sim\mathcal{D}}\left(yf(x;W_i,a) \le 0\right) \le \epsilon, \quad \text{for} \quad n = \widetilde{\Theta}(1/(\gamma^2\epsilon)).$$

Below we prove Theorem 4.3. For any i and W, define

$$\mathcal{R}_i(W) := \ell\left(y_i\left\langle \nabla f_i(W_i), W\right\rangle\right), \quad \text{and} \quad \mathcal{Q}_i(W) := -\ell'\left(y_i\left\langle \nabla f_i(W_i), W\right\rangle\right).$$

Due to homogeneity, it holds that $\mathcal{R}_i(W_i) = \ell\left(y_i f_i(W_i)\right)$ and $\mathcal{Q}_i(W_i) = -\ell'\left(y_i f_i(W_i)\right)$.

The first step is an extension of Lemma 4.5 to the SGD setting, with a similar proof.

Lemma 4.9. With a constant step size $\eta \le 1$, for any $\overline{W}$ and any $i \ge 0$,

$$\eta\left(\sum_{t<i}\mathcal{R}_t(W_t)\right) + \left\|W_i - \overline{W}\right\|_F^2 \le \left\|W_0 - \overline{W}\right\|_F^2 + 2\eta\left(\sum_{t<i}\mathcal{R}_t\left(\overline{W}\right)\right).$$

Proof. Recall that $\left\|\nabla f_t(W_t)\right\|_F \le 1$, we have

$$\left\|W_{t+1} - \overline{W}\right\|_F^2 \le \left\|W_t - \overline{W}\right\|_F^2 - 2\eta\ell'\left(y_t f_t(W_t)\right) y_t\left\langle \nabla f_t(W_t), W_t - \overline{W}\right\rangle + \eta^2\left(\ell'\left(y_t f_t(W_t)\right)\right)^2. \tag{4.12}$$

Similar to the proof of Lemma 4.5, the first order term of eq. (4.12) can be handled using the convexity of ℓ and homogeneity of ReLU as follows

$$\ell'\left(y_t f_t(W_t)\right) y_t\left\langle \nabla f_t(W_t), W_t - \overline{W}\right\rangle \ge \mathcal{R}_t(W_t) - \mathcal{R}_t\left(\overline{W}\right), \tag{4.13}$$

and the second-order term of eq. (4.12) can be bounded as follows

$$\eta^2\left(\ell'\left(y_t f_t(W_t)\right)\right)^2 \le -\eta\ell'\left(y_t f_t(W_t)\right) \le \eta\ell\left(y_t f_t(W_t)\right) = \eta\mathcal{R}_t(W_t), \tag{4.14}$$

since $\eta, -\ell' \le 1$ and $-\ell' \le \ell$. Combining eqs. (4.12) to (4.14) gives

$$\eta\mathcal{R}_t(W_t) \le \left\|W_t - \overline{W}\right\|_F^2 - \left\|W_{t+1} - \overline{W}\right\|_F^2 + 2\eta\mathcal{R}_t\left(\overline{W}\right).$$

Telescoping gives the claim. QED.

With Lemma 4.9, we can also extend Theorem 4.1 to the SGD setting and get a bound on $\sum_{i<n} \mathcal{Q}_i(W_i)$, using a similar proof. With Lemma 4.9, we give the following result, which is an extension of Theorem 4.1 to the SGD setting.

Lemma 4.10. Under Assumption 4.2, given any $\epsilon \in (0,1)$, any $\delta \in (0,1/3)$, and any positive integer n_0, let

$$\lambda := \frac{\sqrt{2\ln(4n_0/\delta)} + \ln(4/\epsilon)}{\gamma/4}, \quad \text{and} \quad M := \frac{4096\lambda^2}{\gamma^6}.$$

For any $m \geq M$ and any constant step size $\eta \leq 1$, if $n_0 \geq n := \lceil 2\lambda^2/(\eta\epsilon) \rceil$, then with probability $1 - 3\delta$,

$$\frac{1}{n}\sum_{i<n} \mathcal{Q}_i(W_i) \leq \epsilon.$$

Proof. We first sample n_0 data examples $(x_0, y_0), \ldots, (x_{n_0-1}, y_{n_0-1})$, and then feed (x_i, y_i) to SGD at step i. We only consider the first n_0 steps.

The proof is similar to the proof of Theorem 4.1. Let n_1 denote the first step before n_0 such that there exists some $1 \leq s \leq m$ with $\left\|w_{s,n_1} - w_{s,0}\right\|_2 > 4\lambda/(\gamma\sqrt{m})$. If such a step does not exist, let $n_1 = n_0$.

Let $\overline{W} := W_0 + \lambda\overline{U}$, in exactly the same way as in Theorem 4.1, we can show that with probability $1 - 3\delta$, for any $0 \leq i < n_1$,

$$y_i \left\langle \nabla f_i(W_i), \overline{W} \right\rangle \geq \ln\left(\frac{4}{\epsilon}\right), \quad \text{and thus} \quad \mathcal{R}_i\left(\overline{W}\right) \leq \epsilon/4.$$

Now consider $n := \lceil 2\lambda^2/(\eta\epsilon) \rceil$. Using Lemma 4.9, in the same way as the proof of Theorem 4.1 (replacing $\widehat{\mathcal{Q}}(W_\tau)$ with $\mathcal{Q}_i(W_i)$, etc.), we can show that $n \leq n_1$. Then invoking Lemma 4.9 again, we get

$$\frac{1}{n}\sum_{i<n} \mathcal{Q}_i(W_i) \leq \frac{1}{n}\sum_{i<n} \mathcal{R}_i(W_i) \leq \frac{\left\|W_0 - \overline{W}\right\|_F^2}{\eta n} + \frac{2}{n}\sum_{i<n} \mathcal{R}_i\left(\overline{W}\right) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

QED.

To further get a bound on the cumulative population risk $\sum_{i<n} \mathcal{Q}(W_i)$, the key observation is that $\sum_{i<n} \big(\mathcal{Q}(W_i) - \mathcal{Q}_i(W_i)\big)$ is a martingale. Using a martingale Bernstein bound, we prove the following lemma; applying it finishes the proof of Theorem 4.3.

Lemma 4.11. Given any $\delta \in (0,1)$, with probability $1-\delta$,

$$\sum_{t<i} \mathcal{Q}(W_t) \le 4 \sum_{t<i} \mathcal{Q}_t(W_t) + 4 \ln\left(\frac{1}{\delta}\right).$$

To prove Lemma 4.11, we need the following martingale Bernstein bound.

Lemma 4.12. [94, Theorem 1] Let $(M_t, \mathcal{F}_t)_{t\ge 0}$ denote a martingale with $M_0 = 0$ and $\mathcal{F}_0$ be the trivial σ-algebra. Let $(\Delta_t)_{t\ge 1}$ denote the corresponding martingale difference sequence, and let

$$V_t := \sum_{j=1}^{t} \mathbb{E}\left[\Delta_j^2 \middle| \mathcal{F}_{j-1}\right]$$

denote the sequence of conditional variance. If $\Delta_t \le R$ a.s., then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$M_t \le \frac{V_t}{R}(e-2) + R \ln\left(\frac{1}{\delta}\right).$$

Proof of Lemma 4.11. For any $i \ge 0$, let z_i denote (x_i, y_i), and $z_{0,i}$ denote $(z_0, \ldots, z_i)$. Note that the quantity $\sum_{t<i}\big(\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t)\big)$ is a martingale w.r.t. the filtration $\sigma(z_{0,i-1})$. The martingale difference sequence is given by $\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t)$, which satisfies

$$\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\ell'\left(yf(x;W_t,a)\right)\right] + \ell'\left(y_t f(x_t;W_t,a)\right) \le 1, \tag{4.15}$$

since $-1 \le \ell' \le 0$. Moreover, we have

$$\begin{aligned}
&\mathbb{E}\left[\big(\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t)\big)^2 \middle| \sigma(z_{0,t-1})\right] \\
&= \mathcal{Q}(W_t)^2 - 2\mathcal{Q}(W_t)\mathbb{E}\left[\mathcal{Q}_t(W_t)\middle|\sigma(z_{0,t-1})\right] + \mathbb{E}\left[\mathcal{Q}_t(W_t)^2\middle|\sigma(z_{0,t-1})\right] \\
&= -\mathcal{Q}(W_t)^2 + \mathbb{E}\left[\mathcal{Q}_t(W_t)^2\middle|\sigma(z_{0,t-1})\right] \\
&\le \mathbb{E}\left[\mathcal{Q}_t(W_t)^2\middle|\sigma(z_{0,t-1})\right] \\
&\le \mathbb{E}\left[\mathcal{Q}_t(W_t)\middle|\sigma(z_{0,t-1})\right] \\
&= \mathcal{Q}(W_t).
\end{aligned} \tag{4.16}$$

Invoking Lemma 4.12 with eqs. (4.15) and (4.16) gives that with probability $1-\delta$,

$$\sum_{t<i}\big(\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t)\big) \le (e-2)\sum_{t<i}\mathcal{Q}(W_t) + \ln\left(\frac{1}{\delta}\right),$$

which implies $\sum_{t<i}\mathcal{Q}(W_t) \le 4\sum_{t<i}\mathcal{Q}_t(W_t) + 4\ln\left(\frac{1}{\delta}\right)$. QED.

Finally, we prove Theorem 4.3.

Proof of Theorem 4.3. Suppose the condition of Lemma 4.10 holds. Then we have for $n = \lceil 2\lambda^2/(\eta\epsilon)\rceil$, with probability $1-3\delta$,

$$\frac{1}{n}\sum_{i<n}\mathcal{Q}_i(W_i) \le \epsilon.$$

Further invoking Lemma 4.11 gives that with probability $1-4\delta$,

$$\frac{1}{n}\sum_{i<n}\mathcal{Q}(W_i) \le \frac{4}{n}\sum_{i<n}\mathcal{Q}_i(W_i) + \frac{4}{n}\ln\left(\frac{1}{\delta}\right) \le 5\epsilon.$$

Since $P_{(x,y)\sim\mathcal{D}}\left(yf(x;W,a)\le 0\right) \le 2\mathcal{Q}(W)$, we get

$$\frac{1}{n}\sum_{i=1}^{n} P_{(x,y)\sim\mathcal{D}}\left(yf(x;W_i,a)\le 0\right) \le 10\epsilon.$$

For the condition of Lemma 4.10 to hold, it is enough to let

$$n_0 = \Theta\left(\frac{\ln(1/\delta)}{\eta\gamma^2\epsilon^2}\right),$$

which gives

$$M = \Theta\left(\frac{\ln(1/\delta)+\ln(1/\epsilon)^2}{\gamma^8}\right) \quad \text{and} \quad n = \Theta\left(\frac{\ln(1/\delta)+\ln(1/\epsilon)^2}{\gamma^2\epsilon}\right).$$

QED.

4.4 ON SEPARABILITY

In this section we give some discussion on Assumption 4.1, the separability of the NTK.

Given a training set $\left\{(x_i,y_i)\right\}_{i=1}^n$, the linear kernel is defined as $K_0(x_i,x_j) := \left\langle x_i, x_j\right\rangle$. The maximum margin achievable by a linear classifier is given by

$$\gamma_0 := \min_{q\in\Delta_n} \sqrt{(q\odot y)^\top K_0\,(q\odot y)}. \tag{4.17}$$

where Δ_n denotes the probability simplex and $\odot$ denotes the Hadamard product. In addition to the dual definition eq. (4.17), when $\gamma_0 > 0$ there also exists a maximum margin classifier $\bar{u}$ which gives a primal characterization of γ_0: it holds that $\|\bar{u}\|_2 = 1$ and $y_i\left\langle \bar{u}, x_i\right\rangle \ge \gamma_0$ for

all i.

In this paper we consider another kernel, the infinite-width NTK with respect to the first layer:

$$\begin{aligned} K_1\left(x_i, x_j\right) &:= \mathbb{E}\left[\frac{\partial f(x_i; W_0, a)}{\partial W_0}, \frac{\partial f(x_j; W_0, a)}{\partial W_0}\right] \\ &= \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\left\langle x_i \mathbb{1}\left[\langle x_i, w\rangle > 0\right], x_j \mathbb{1}\left[\langle x_j, w\rangle > 0\right]\right\rangle\right] = \langle \phi_i, \phi_j\rangle_{\mathcal{H}}. \end{aligned}$$

Here ϕ and $\mathcal{H}$ are defined at the beginning of Section 4.1. Similar to the dual definition of γ_0, the margin given by K_1 is defined as

$$\gamma_1 := \min_{q\in\Delta_n} \sqrt{(q\odot y)^\top K_1 (q\odot y)}. \tag{4.18}$$

We can also give a primal characterization of γ_1 when it is positive; the proof uses the Fenchel duality theory.

Proposition 4.4. If $\gamma_1 > 0$, then there exists $\hat{v} \in \mathcal{H}$ such that $\|\hat{v}\|_{\mathcal{H}} = 1$, and $y_i \langle \hat{v}, \phi_i\rangle_{\mathcal{H}} \geq \gamma_1$ for any $1 \leq i \leq n$. Additionally $\left\|\hat{v}(z)\right\|_2 \leq 1/\gamma_1$ for any $z \in \mathbb{R}^d$.

Proof. Define $f : \mathcal{H} \to \mathbb{R}$ by

$$f(w) := \frac{1}{2}\int \|w(z)\|_2^2 \,\mathrm{d}\mu_{\mathcal{N}}(z) = \frac{1}{2}\|w\|_{\mathcal{H}}^2.$$

It holds that f is continuous, and f^* has the same form. Define $g : \mathbb{R}^n \to \mathbb{R}$ by

$$g(p) := \max_{1\leq i\leq n} p_i,$$

with conjugate

$$g^*(q) = \begin{cases} 0, & \text{if } q \in \Delta_n, \\ +\infty, & \text{o.w.} \end{cases}$$

Finally, define the linear mapping $A : \mathcal{H} \to \mathbb{R}^n$ by $(Aw)_i = y_i \langle w, \phi_i\rangle_{\mathcal{H}}$.

Since f, f^*, g and g^* are lower semi-continuous, and $\mathbf{dom}\, g - A\mathbf{dom}\, f = \mathbb{R}^n$, and $\mathbf{dom}\, f^* - A^*\mathbf{dom}\, g^* = \mathcal{H}$, Fenchel duality may be applied in each direction [95, Theorem 4.4.3], and ensures that

$$\inf_{w\in\mathcal{H}} \left(f(w) + g(Aw)\right) = \sup_{q\in\mathbb{R}^n} \left(-f^*(A^*q) - g^*(-q)\right).$$

with optimal primal-dual solutions $(\bar{w}, \bar{q})$. Moreover

$$\begin{aligned}
\inf_{w\in\mathcal{H}} \big(f(w) + g(Aw)\big) &= \inf_{w\in\mathcal{H}, u\in\mathbb{R}^n} \sup_{q\in\mathbb{R}^n} \big(f(w) + g(Aw+u) + \langle q, u\rangle\big) \\
&\geq \sup_{q\in\mathbb{R}^n} \inf_{w\in\mathcal{H}, u\in\mathbb{R}^n} \big(f(w) + g(Aw+u) + \langle q, u\rangle\big) \\
&= \sup_{q\in\mathbb{R}^n} \inf_{w\in\mathcal{H}, u\in\mathbb{R}^n} \Big(\big(f(w) - \langle A^*q, w\rangle\big)_{\mathcal{H}} + \big(g(Aw+u) - \langle -q, Aw+u\rangle\big)\Big) \\
&= \sup_{q\in\mathbb{R}^n} \big(-f^*(A^*q) - g^*(-q)\big).
\end{aligned}$$

By strong duality, the inequality holds with equality. It follows that

$$\bar{w} = A^*\bar{q}, \quad \text{and} \quad \mathbf{supp}(-\bar{q}) \subset \arg\max_{1\leq i\leq n} (A\bar{w})_i.$$

Now let us look at the dual optimization problem. It is clear that

$$\sup_{q\in\mathbb{R}^n} \big(-f^*(A^*q) - g^*(-q)\big) = -\inf_{q\in\Delta_n} f^*(A^*q).$$

In addition, we have

$$\begin{aligned}
f^*(A^*q) &= \frac{1}{2}\int \left\| \sum_{i=1}^n q_i y_i \phi_i(z) \right\|_2^2 \mathrm{d}\mu_{\mathcal{N}}(z) \\
&= \frac{1}{2}\int \sum_{i,j=1}^n q_i q_j y_i y_j \left\langle \phi_i(z), \phi_j(z)\right\rangle \mathrm{d}\mu_{\mathcal{N}}(z) \\
&= \frac{1}{2}\sum_{i,j=1}^n q_i q_j y_i y_j \int \left\langle \phi_i(z), \phi_j(z)\right\rangle \mathrm{d}\mu_{\mathcal{N}}(z) \\
&= \frac{1}{2}\sum_{i,j=1}^n q_i q_j y_i y_j K_1(i,j) = \frac{1}{2}(q\odot y)^\top K_1 (q\odot y),
\end{aligned}$$

and thus $f^*(A^*\bar{q}) = \gamma_1^2/2$. Since $\bar{w} = A^*\bar{q}$, we have that $\|\bar{w}\|_{\mathcal{H}} = \gamma_1$. In addition,

$$g(A\bar{w}) = -f^*\left(A^*\bar{q}\right) - f\left(\bar{w}\right) = -\gamma_1^2,$$

and thus $-\bar{w}$ has margin γ_1^2. Moreover, we have

$$\bar{w}(z) = \sum_{i=1}^n \bar{q}_i y_i \phi_i(z) = \sum_{i=1}^n \bar{q}_i y_i x_i \mathbb{1}\left[\langle z, x_i\rangle > 0\right],$$

and thus $\left\|\bar{w}(z)\right\|_2 \leq 1$. Therefore, $\hat{v} = -\bar{w}/\gamma_1$ satisfies all requirements of Proposition 4.4. QED.

Using the upper bound $\left\|\hat{v}(z)\right\|_2 \leq 1/\gamma_1$, we can see that $\gamma_1\hat{v}$ satisfies Assumption 4.1 with $\gamma \geq \gamma_1^2$. However, such an upper bound $\left\|\hat{v}(z)\right\|_2 \leq 1/\gamma_1$ might be too loose, which leads to a bad rate. In fact, as shown later, in some cases we can construct $\bar{v}$ directly which satisfies Assumption 4.1 with a large γ. For this reason, we choose to make Assumption 4.1 instead of assuming a positive γ_1.

However, we can use γ_1 to show that Assumption 4.1 always holds when there are no parallel inputs. [77, Corollary I.2] proved that if for any two feature vectors x_i and x_j, we have $\|x_i - x_j\|_2 \geq \theta$ and $\|x_i + x_j\|_2 \geq \theta$ for some $\theta > 0$, then the minimum eigenvalue of K_1 is at least $\theta/(100n^2)$. For arbitrary labels $y \in \{-1, +1\}^n$, since $\|q \odot y\|_2 \geq 1/\sqrt{n}$, we have the worst case bound $\gamma_1^2 \geq \theta/(100n^3)$. A direct improvement of this bound is $\theta/(100n_S^3)$, where n_S denotes the number of support vectors, which could be much smaller than n with real world data.

On the other hand, given any training set $\left\{(x_i, y_i)\right\}_{i=1}^n$ which may have a large margin, replacing y with random labels would destroy the margin, which is what should be expected.

Proposition 4.5. Given any training set $\left\{(x_i, y_i)\right\}_{i=1}^n$, if the true labels y are replaced with random labels $\epsilon \sim \text{unif}\left(\{-1, +1\}^n\right)$, then with probability 0.9 over the random labels, it holds that $\gamma_1 \leq 1/\sqrt{20n}$.

Proof. Let $\hat{q}$ denote the uniform probability vector $(1/n, \ldots, 1/n)$. Note that

$$\begin{aligned}
\mathbb{E}_{\epsilon\sim\text{unif}\left(\{-1,+1\}^n\right)}\left[(\hat{q}\odot\epsilon)^\top K_1 (\hat{q}\odot\epsilon)\right] &= \mathbb{E}_{\epsilon\sim\text{unif}\left(\{-1,+1\}^n\right)}\left[\sum_{i,j=1}^n \frac{1}{n^2}\epsilon_i\epsilon_j K_1(x_i, x_j)\right] \\
&= \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}_{\epsilon\sim\text{unif}\left(\{-1,+1\}^n\right)}\left[\epsilon_i\epsilon_j K_1(x_i, x_j)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n K_1(x_i, x_i) = \frac{1}{2n}.
\end{aligned}$$

Since $0 \leq (\hat{q}\odot\epsilon)^\top K_1 (\hat{q}\odot\epsilon) \leq 1$ for any ϵ, by Markov's inequality with probability 0.9, it holds that $(\hat{q}\odot\epsilon)^\top K_1 (\hat{q}\odot\epsilon) \leq 1/(20n)$, and thus $\gamma_1 \leq 1/\sqrt{20n}$. QED.

Although the above bounds all have a polynomial dependency on n, they hold for arbitrary or random labels, and thus do not assume any relationship between the features and labels.

Next we give some examples where there is a strong feature-label relationship, and thus a much larger margin can be proved.

4.4.1 The linearly separable case

Suppose the data distribution is linearly separable with margin γ_0: there exists a unit vector $\bar{u}$ such that $y \langle \bar{u}, x \rangle \geq \gamma_0$ almost surely. Then we can define $\bar{v}(z) := \bar{u}$ for any $z \in \mathbb{R}^d$. For almost all (x, y), we have

$$\begin{aligned} y \int \langle \bar{v}(z), x \rangle \mathbb{1}\left[\langle z, x \rangle > 0\right] \mathrm{d}\mu_{\mathcal{N}}(z) &= \int y \langle \bar{u}, x \rangle \mathbb{1}\left[\langle z, x \rangle > 0\right] \mathrm{d}\mu_{\mathcal{N}}(z) \\ &\geq \gamma \int \mathbb{1}\left[\langle z, x \rangle > 0\right] \mathrm{d}\mu_{\mathcal{N}}(z) \\ &= \frac{\gamma_0}{2}, \end{aligned}$$

and thus Assumption 4.1 holds with $\gamma = \gamma_0/2$.

4.4.2 The noisy 2-XOR distribution

We consider the noisy 2-XOR distribution introduced in [74]. It is the uniform distribution over the following 2^d points:

$$\begin{aligned} &(x_1, x_2, y, x_3, \ldots, x_d) \\ \in &\left\{ \left(\frac{1}{\sqrt{d-1}}, 0, 1 \right), \left(0, \frac{1}{\sqrt{d-1}}, -1 \right), \left(\frac{-1}{\sqrt{d-1}}, 0, 1 \right), \left(0, \frac{-1}{\sqrt{d-1}}, -1 \right) \right\} \\ &\times \left\{ \frac{-1}{\sqrt{d-1}}, \frac{1}{\sqrt{d-1}} \right\}^{d-2}. \end{aligned}$$

The factor $1/\sqrt{d-1}$ ensures that $\|x\|_2 = 1$, and $\times$ above denotes the Cartesian product. Here the label y only depends on the first two coordinates of the input x.

To construct $\bar{v}$, we first decompose $\mathbb{R}^2$ into four regions:

$$\begin{aligned} A_1 &:= \left\{ (z_1, z_2) \;\middle|\; z_1 \geq 0, |z_1| \geq |z_2| \right\}, \\ A_2 &:= \left\{ (z_1, z_2) \;\middle|\; z_2 > 0, |z_1| < |z_2| \right\}, \\ A_3 &:= \left\{ (z_1, z_2) \;\middle|\; z_1 \leq 0, |z_1| \geq |z_2| \right\} \setminus \{(0,0)\}, \\ A_4 &:= \left\{ (z_1, z_2) \;\middle|\; z_2 < 0, |z_1| < |z_2| \right\}. \end{aligned}$$

Then $\bar{v}$ can de defined as follows: (i) for $(z_1, z_2) \in A_1$, let $\bar{v} := (1, 0, 0, \ldots, 0)$; (ii) for $(z_1, z_2) \in A_2$, let $\bar{v} := (0, -1, 0, \ldots, 0)$; (iii) for $(z_1, z_2) \in A_3$, let $\bar{v} := (-1, 0, 0, \ldots, 0)$; (iv) for $(z_1, z_2) \in A_4$, let $\bar{v} := (0, 1, 0, \ldots, 0)$.

The following result shows that $\gamma = \Omega(1/d)$. Note that n could be as large as 2^d, in which case γ is basically $O\left(1/\ln(n)\right)$.

Proposition 4.6. For any (x, y) sampled from the noisy 2-XOR distribution and any $d \geq 3$,

$$y \int \left\langle \bar{v}(z), x \right\rangle \mathbb{1}\left[\langle z, x\rangle > 0\right] \mathrm{d}\mu_{\mathcal{N}}(z) \geq \frac{1}{60d}.$$

Proof. By symmetry, we only need to consider an (x, y) where $(x_1, x_2, y) = (1/\sqrt{d-1}, 0, 1)$. Let $z_{p,q}$ denote $(z_p, z_{p+1}, \ldots, z_q)$, and similarly define $x_{p,q}$. We have

$$\begin{aligned}
& y \int \left\langle \bar{v}(z), x \right\rangle \mathbb{1}\left[\langle z, x\rangle > 0\right] \mathrm{d}\mu_{\mathcal{N}}(z) \\
&= y \int \left(\int \left\langle \bar{v}(z), x \right\rangle \mathbb{1}\left[\langle z, x\rangle > 0\right] \mathrm{d}\mu_{\mathcal{N}}(z_{3,d}) \right) \mathrm{d}\mu_{\mathcal{N}}(z_{1,2}) && (4.19) \\
&= y \int \left\langle \bar{v}(z)_{1,2}, x_{1,2} \right\rangle \left(\int \mathbb{1}\left[\left\langle z_{1,2}, x_{1,2} \right\rangle + \left\langle z_{3,d}, x_{3,d} \right\rangle > 0 \right] \mathrm{d}\mu_{\mathcal{N}}(z_{3,d}) \right) \mathrm{d}\mu_{\mathcal{N}}(z_{1,2}) && (4.20) \\
&= \sum_{i=1}^{4} y \int \left\langle \bar{v}(z)_{1,2}, x_{1,2} \right\rangle \Big(\int \mathbb{1}\left[\left\langle z_{1,2}, x_{1,2} \right\rangle + \left\langle z_{3,d}, x_{3,d} \right\rangle > 0 \right] \mathrm{d}\mu_{\mathcal{N}}(z_{3,d}) \Big) \mathbb{1}\left[z_{1,2} \in A_i\right] \mathrm{d}\mu_{\mathcal{N}}(z_{1,2}), && (4.21)
\end{aligned}$$

where eq. (4.19) is due to the independence between $z_{1,2}$ and $z_{3,d}$, and in eq. (4.20) we use the fact that $\bar{v}(z)_{1,2}$ only depends on $z_{1,2}$ and $\bar{v}(z)_{3,d}$ are all zero. Since $\left\langle \bar{v}(z)_{1,2}, x_{1,2} \right\rangle = 0$ for $z_{1,2} \in A_2 \cup A_4$, we only need to consider A_1 and A_3 in eq. (4.21). For simplicity, we will denote $z_{1,2}$ by $p \in \mathbb{R}^2$, and $\bar{v}(z)_{1,2}$ by $\bar{v}(p)$, and $z_{3,d}$ by $q \in \mathbb{R}^{d-2}$.

For any nonzero $p \in A_1$, we have $-p \in A_3$, and $\left\langle \bar{v}(p), x_{1,2} \right\rangle = 1/\sqrt{d-1}$. Therefore

$$\begin{aligned}
& y \left\langle \bar{v}(p), x_{1,2} \right\rangle \left(\int \mathbb{1}\left[\left\langle p, x_{1,2} \right\rangle + \left\langle q, x_{3,d} \right\rangle > 0 \right] \mathrm{d}\mu_{\mathcal{N}}(q) \right) \\
& \quad + y \left\langle \bar{v}(-p), x_{1,2} \right\rangle \left(\int \mathbb{1}\left[\left\langle -p, x_{1,2} \right\rangle + \left\langle q, x_{3,d} \right\rangle > 0 \right] \mathrm{d}\mu_{\mathcal{N}}(q) \right) \\
&= \frac{1}{\sqrt{d-1}} \int \left(\mathbb{1}\left[\frac{p_1}{\sqrt{d-1}} + \left\langle q, x_{3,d} \right\rangle > 0 \right] - \mathbb{1}\left[\frac{-p_1}{\sqrt{d-1}} + \left\langle q, x_{3,d} \right\rangle > 0 \right] \right) \mathrm{d}\mu_{\mathcal{N}}(q) \\
&= \frac{1}{\sqrt{d-1}} \mathbb{P}\left(\frac{-p_1}{\sqrt{d-1}} \leq \left\langle q, x_{3,d} \right\rangle \leq \frac{p_1}{\sqrt{d-1}} \right). && (4.22)
\end{aligned}$$

Let φ denote the density function of the standard Gaussian distribution, and for $c > 0$, let

$U(c)$ denote the probability that a standard Gaussian random variable lies in the interval $[-c, c]$:

$$U(c) := \int_{-c}^{c} \varphi(t)\,\mathrm{d}t.$$

Since $\langle q, x_{3,d}\rangle$ is a Gaussian variable with standard deviation $\sqrt{(d-2)/(d-1)}$, we have

$$\mathbb{P}\left(\frac{-p_1}{\sqrt{d-1}} \le \langle q, x_{3,d}\rangle \le \frac{p_1}{\sqrt{d-1}}\right) = U\left(\frac{p_1}{\sqrt{d-2}}\right). \tag{4.23}$$

Plugging eqs. (4.22) and (4.23) into eq. (4.21) gives:

$$\begin{aligned}
y\int \langle \bar{v}(z), x\rangle\, \mathbb{1}\left[\langle z, x\rangle > 0\right] \mathrm{d}\mu_{\mathcal{N}}(z) &= \frac{1}{\sqrt{d-1}} \int U\left(\frac{p_1}{\sqrt{d-2}}\right) \mathbb{1}\left[p \in A_1\right] \mathrm{d}\mu_{\mathcal{N}}(p) \\
&= \frac{1}{\sqrt{d-1}} \int_0^{\infty} U\left(\frac{p_1}{\sqrt{d-2}}\right) \left(\int_{-p_1}^{p_1} \varphi(p_2)\,\mathrm{d}p_2\right) \varphi(p_1)\,\mathrm{d}p_1 \\
&= \frac{1}{\sqrt{d-1}} \int_0^{\infty} U\left(\frac{p_1}{\sqrt{d-2}}\right) U(p_1)\varphi(p_1)\,\mathrm{d}p_1 \\
&\ge \frac{1}{\sqrt{d-1}} \int_0^{1} U\left(\frac{p_1}{\sqrt{d-2}}\right) U(p_1)\varphi(p_1)\,\mathrm{d}p_1.
\end{aligned}$$

For $t \in [-1, +1]$, it holds that $\varphi(t) \ge 1\sqrt{2\pi e}$, and thus

$$U(a) = \int_{-a}^{a} \varphi(t)\,\mathrm{d}t \ge \frac{2a}{\sqrt{2\pi e}}.$$

Therefore eq. (4.21) is lower bounded by

$$\begin{aligned}
\frac{1}{\sqrt{d-1}} \int_0^{1} U\left(\frac{p_1}{\sqrt{d-2}}\right) U(p_1)\varphi(p_1)\,\mathrm{d}p_1 &\ge \frac{1}{\sqrt{d-1}} \int_0^{1} \frac{2}{\sqrt{2\pi e}} \cdot \frac{p_1}{\sqrt{d-2}} \cdot \frac{2p_1}{\sqrt{2\pi e}} \cdot \frac{1}{\sqrt{2\pi e}}\,\mathrm{d}p_1 \\
&\ge \frac{1}{20\sqrt{(d-1)(d-2)}} \int_0^{1} p_1^2\,\mathrm{d}p_1 \\
&= \frac{1}{60\sqrt{(d-1)(d-2)}} \\
&\ge \frac{1}{60d}.
\end{aligned}$$

QED.

We can prove two other interesting results for the noisy 2-XOR data.

The width needs a $\mathrm{poly}(1/\gamma)$ **dependency for initial separability.** The first step of an NTK analysis is to show that the gradient features at initialization $\left\{\left(\nabla f_i(W_0), y_i\right)\right\}_{i=1}^n$ is separable. Proposition 4.7 gives an example where $\left\{\left(\nabla f_i(W_0), y_i\right)\right\}_{i=1}^n$ is nonseparable when the network is narrow.

Proposition 4.7. Let $D = \{(x_i, y_i)\}_{i=1}^4$ denote an arbitrary subset of the noisy 2-XOR dataset such that x_i's have the same last $(d-2)$ coordinates. For any $d \geq 20$, if $m \leq \sqrt{d-2}/4$, then with probability $1/2$ over the random initialization of W_0, for any weights $V \in \mathbb{R}^{m \times d}$, it holds that $y_i \left\langle V, \nabla f_i(W_0) \right\rangle \leq 0$ for at least one $i \in \{1, 2, 3, 4\}$.

For the noisy 2-XOR data, the separator $\bar{v}$ has margin $\gamma = \Omega(1/d)$, and $1/\gamma = O(d)$. As a result, if we want $\left\{\left(\nabla f_i(W_0), y_i\right)\right\}_{i=1}^n$ to be separable, the width has to be $\Omega(1/\sqrt{\gamma})$. For a smaller width, gradient descent might still be able to solve the problem, but a beyond-NTK analysis would be needed.

To prove Proposition 4.7, we need the following technical lemma.

Lemma 4.13. Given $z_1 \sim \mathcal{N}(0,1)$ and $z_2 \sim \mathcal{N}(0, b^2)$ that are independent where $b > 1$, we have

$$\mathbb{P}\left(|z_1| < |z_2|\right) > 1 - \frac{1}{b}.$$

Proof. First note that for $z_3 \sim \mathcal{N}(0,1)$ which is independent of z_1,

$$\mathbb{P}\left(|z_1| < |z_2|\right) = \mathbb{P}\left(|z_1| < b|z_3|\right) = 1 - \mathbb{P}\left(|z_3| < \frac{1}{b}|z_1|\right).$$

Still let φ denote the density of $\mathcal{N}(0,1)$, and let $U(c)$ denote the probability that $z_3 \in [-c, c]$. We have

$$\begin{aligned}
\mathbb{P}\left(|z_3| < \frac{1}{b}|z_1|\right) &= \int\int \mathbb{1}\left[|z_3| < \frac{1}{b}|z_1|\right] \varphi(z_3)\varphi(z_1)\,\mathrm{d}z_3\,\mathrm{d}z_1 \\
&= \int U\left(\frac{1}{b}|z_1|\right)\varphi(z_1)\,\mathrm{d}z_1 \\
&\leq \frac{2}{\sqrt{2\pi}b}\int |z_1|\varphi(z_1)\,\mathrm{d}z_1 = \frac{2}{\pi b} < \frac{1}{b},
\end{aligned}$$

where we use the facts that $U(c) \leq 2c/\sqrt{2\pi}$ and $\mathbb{E}[|z_1|] = \sqrt{2/\pi}$. QED.

We now give the proof of Proposition 4.7 using Lemma 4.13.

Proof of Proposition 4.7. By symmetry, we only need to consider the following training set:

$$\begin{aligned} x_1 &= (1, 0, 1, \ldots, 1), \quad y_1 = 1, \\ x_2 &= (0, 1, 1, \ldots, 1), \quad y_2 = -1, \\ x_3 &= (-1, 0, 1, \ldots, 1), \quad y_3 = 1, \\ x_4 &= (0, -1, 1, \ldots, 1), \quad y_4 = -1. \end{aligned}$$

The $1/\sqrt{d-1}$ factor is omitted also because we only discuss the 0/1 loss.

For any s, let A_s denote the event that

$$\mathbb{1}\left[\langle w_s, x_1\rangle > 0\right] = \mathbb{1}\left[\langle w_s, x_2\rangle > 0\right] = \mathbb{1}\left[\langle w_s, x_3\rangle > 0\right] = \mathbb{1}\left[\langle w_s, x_4\rangle > 0\right].$$

We will show that if $m \le \sqrt{d-2}/4$, then A_s is true for all $1 \le s \le m$ with probability $1/2$, and Proposition 4.7 follows from the fact that the XOR data is not linearly separable.

For any s and i,

$$\langle w_s, x_i\rangle = (w_s)_1(x_i)_1 + (w_s)_2(x_i)_2 + \sum_{j=3}^{d}(w_s)_j.$$

Since $\big((x_i)_1, (x_i)_2\big)$ is $(1,0)$ or $(0,1)$ or $(-1,0)$ or $(0,-1)$, event A_s will happen as long as

$$\big|(w_s)_1\big| < \left|\sum_{j=3}^{d}(w_s)_j\right|, \quad \text{and} \quad \big|(w_s)_2\big| < \left|\sum_{s=3}^{d}(w_s)_j\right|.$$

Note that $(w_s)_1, (w_s)_2 \sim \mathcal{N}(0,1)$ while $\sum_{j=3}^{d}(w_s)_j \sim \mathcal{N}(0, d-2)$. As a result, due to Lemma 4.13,

$$\mathbb{P}\left(\big|(w_s)_1\big| < \left|\sum_{j=3}^{d}(w_s)_j\right|\right) = \mathbb{P}\left(\big|(w_s)_2\big| < \left|\sum_{s=3}^{d}(w_s)_j\right|\right) > 1 - \frac{1}{\sqrt{d-2}}.$$

Using a union bound, $\mathbb{P}(A_s) > 1 - 2/\sqrt{d-2}$. If $m \le \sqrt{d-2}/4$, then by a union bound again,

$$\mathbb{P}\left(\bigcup_{1\le s\le m} A_s\right) > 1 - \frac{2}{\sqrt{d-2}}m \ge 1 - \frac{2}{\sqrt{d-2}}\frac{\sqrt{d-2}}{4} = \frac{1}{2}.$$

QED.

A tight sample complexity upper bound for the infinite-width NTK. [74] give a d^2 sample complexity lower bound for any NTK classifier on the noisy 2-XOR data. It turns out that γ could give a *matching* sample complexity upper bound for the NTK and SGD.

[74] consider the infinite-width NTK with respect to both layers. For the first layer, the infinite-width NTK K_1 is defined in Section 4.4, and the corresponding RKHS $\mathcal{H}$ and RKHS mapping ϕ is defined in Section 4.1. For the second layer, the infinite width NTK is defined by

$$\begin{aligned} K_2\left(x_i, x_j\right) &:= \mathbb{E}\left[\frac{\partial f(x_i; W_0, a)}{\partial a}, \frac{\partial f(x_j; W_0, a)}{\partial a}\right] \\ &= \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\sigma\left(\langle w, x_i\rangle\right)\sigma\left(\langle w, x_j\rangle\right)\right]. \end{aligned}$$

The corresponding RKHS $\mathcal{K}$ and inner product $\langle w_1, w_2\rangle_{\mathcal{K}}$ are given by

$$\mathcal{K} := \left\{ w : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \int w(z)^2 \,\mathrm{d}\mu_{\mathcal{N}}(z) < \infty \right\}, \quad \text{and} \quad \langle w_1, w_2\rangle_{\mathcal{K}} = \int w_1(z) w_2(z) \,\mathrm{d}\mu_{\mathcal{N}}(z).$$

Given any $x \in \mathbb{R}^d$, it is mapped into $\psi_x \in \mathcal{K}$, where $\psi_x(z) := \sigma\left(\langle z, x\rangle\right)$. It holds that $K_2(x_i, x_j) = \langle \psi_{x_i}, \psi_{x_j}\rangle_{\mathcal{K}}$. The infinite-width NTK with respect to both layers is just $K_1 + K_2$. The corresponding RHKS is just $\mathcal{H} \times \mathcal{K}$ with the inner product

$$\langle (v_1, w_1), (v_2, w_2)\rangle_{\mathcal{H}\times\mathcal{K}} = \langle v_1, v_2\rangle_{\mathcal{H}} + \langle w_1, w_2\rangle_{\mathcal{K}}.$$

The classifier $\bar{v}$ constructed earlier has a unit norm (i.e., $\|\bar{v}\|_{\mathcal{H}} = 1$) and margin γ on the space $\mathcal{H}$. On $\mathcal{H} \times \mathcal{K}$, it is enough to consider $(\bar{v}, 0)$, which also has a unit norm and margin γ. Since the infinite-width NTK model is a linear model in $\mathcal{H} \times \mathcal{K}$, Theorem 2.1 can be used to show that SGD on the RKHS $\mathcal{H} \times \mathcal{K}$ could obtain a test error of ϵ with a sample complexity of $\widetilde{O}(1/(\gamma^2\epsilon))$. (The analysis of Theorem 2.1 is done in $\mathbb{R}^d$, but it still works with a well-defined inner product.) Since $\gamma = \Omega(1/d)$, to achieve a constant test accuracy we need $\widetilde{O}(d^2)$ samples. This matches (up to logarithmic factors) the sample complexity lower bound of d^2 given by [74].

Chapter 5: Deep homogeneous networks

In this chapter, we analyze the implicit bias of gradient flow on deep networks. We focus on gradient flow to illustrate the key proof ideas, but many results can be extended to gradient descent with small enough learning rates.

First, in Section 5.1, we consider deep linear networks, which maps the input x to $W_L W_{L-1} \cdots W_2 W_1 x$, where $W_1, \ldots, W_L$ denote L weight matrices (layers). A deep linear network can still only represent a linear function, but it induces a nonconvex optimization problem, which makes the analysis much trickier. We will show that, despite overparameterization and nonconvexity, under some mild conditions, gradient flow still learns a simple solution: all weight matrices become nearly rank-1, adjacent weight matrices tend to have identical top singular vectors, and the whole network computes the maximum-margin predictor (cf. Theorems 5.1 and 5.2).

In Section 5.2, we further extend the previous results to deep homogeneous networks. We show that the gradient flow iterate w_t and the corresponding (negative) gradient $-\nabla\widehat{\mathcal{R}}(w_t)$ converge to the same direction (cf. Theorem 5.3); this result implies the previous result for deep linear networks, and can also be applied in many other settings.

5.1 ALIGNMENT IN DEEP LINEAR NETWORKS

In this section, as in Chapter 2, we consider a training set that is linearly separable, and let γ^* and u^* denote the maximum margin and maximum-margin predictor.

A linear network of depth L is parameterized by weight matrices $W_L, \ldots, W_1$, where $W_k \in \mathbb{R}^{d_k \times d_{k-1}}$, $d_0 = d$, and $d_L = 1$. Let $W = (W_L, \ldots, W_1)$ denote all parameters of the network. The (empirical) risk induced by the network is given by

$$\widehat{\mathcal{R}}(W) = \widehat{\mathcal{R}}\left(W_L, \ldots, W_1\right) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i W_L \cdots W_1 x_i\right) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\langle w_{\text{prod}}, z_i\rangle\right),$$

where $w_{\text{prod}} := (W_L \cdots W_1)^\top$, and $z_i := y_i x_i$.

In this section, we consider loss functions satisfying the following conditions.

Assumption 5.1. $\ell' < 0$ is continuous, $\lim_{x\to-\infty} \ell(x) = \infty$ and $\lim_{x\to\infty} \ell(x) = 0$.

Moreover, we consider gradient flow $\left\{W(t) \middle| t \geq 0, t \in \mathbb{R}\right\}$, which starts from some $W(0)$ at $t = 0$, and proceeds as

$$\frac{\mathrm{d}W(t)}{\mathrm{d}t} = -\nabla\widehat{\mathcal{R}}\left(W(t)\right).$$

We assume that the initialization of the network is not a critical point and induces a risk no larger than the risk of the trivial linear predictor 0.

Assumption 5.2. The initialization $W(0)$ satisfies $\nabla\widehat{\mathcal{R}}\left(W(0)\right) \neq 0$ and $\widehat{\mathcal{R}}\left(W(0)\right) \leq \widehat{\mathcal{R}}(0) = \ell(0)$.

It is natural to require that the initialization is not a critical point, since otherwise gradient flow will never make a progress. The requirement $\widehat{\mathcal{R}}\left(W(0)\right) \leq \widehat{\mathcal{R}}(0)$ can be easily satisfied, for example, by making $W_1(0) = 0$ and $W_L(0)\cdots W_2(0) \neq 0$. Alternatively, we can ensure Assumption 5.2 using an NTK analysis.

For convenience, only within this section, we will use W, W_k, and w_{prod} to denote $W(t)$, $W_k(t)$ and $w_{\text{prod}}(t)$.

Previously, [96] considered gradient descent on fully connected linear networks and linear convolutional networks. In particular, for the exponential loss, assuming the risk is minimized to 0 and the gradients converge in direction, they showed that the whole network converges in direction to the maximum margin solution. These two assumptions are on the gradient descent process itself; by contrast, in the following we will show alignment and margin maximization results only assuming Assumption 5.2. The analysis is based on [31].

Additional related work. There has been a rich literature on linear networks. [97] analyzed the learning dynamics of deep linear networks, showing that they exhibit some learning patterns similar to nonlinear networks, such as a long plateau followed by a rapid risk drop. [98] showed that depth can help accelerate optimization. On the landscape properties of deep linear networks, [99, 100] showed that under various structural assumptions, all local optima are global. [101] gave a necessary and sufficient characterization of critical points for deep linear networks.

5.1.1 Risk convergence and layer alignment

One key property of gradient flow is that it never increases the risk:

$$\frac{\mathrm{d}\widehat{\mathcal{R}}(W)}{\mathrm{d}t} = \left\langle \nabla\widehat{\mathcal{R}}(W), \frac{\mathrm{d}W}{\mathrm{d}t} \right\rangle = -\|\nabla\widehat{\mathcal{R}}(W)\|_F^2 = -\sum_{k=1}^{L} \left\| \frac{\partial\widehat{\mathcal{R}}}{\partial W_k} \right\|_F^2 \leq 0. \tag{5.1}$$

We now state the main result: under Assumptions 5.1 and 5.2, gradient flow minimizes the risk, W_k and w_{prod} all go to infinity, and the alignment phenomenon occurs.

Theorem 5.1. Under Assumptions 5.1 and 5.2, gradient flow iterates satisfy:

- $\lim_{t\to\infty} \widehat{\mathcal{R}}(W) = 0$.
- For any $1 \le k \le L$, $\lim_{t\to\infty} \|W_k\|_F = \infty$.
- For any $1 \le k \le L$, letting (u_k, v_k) denote the first left and right singular vectors of W_k,
$$\lim_{t\to\infty} \left\| \frac{W_k}{\|W_k\|_F} - u_k v_k^\top \right\|_F = 0.$$
Moreover, for any $1 \le k < L$, $\lim_{t\to\infty} \left|\langle v_{k+1}, u_k \rangle\right| = 1$. As a result,
$$\lim_{t\to\infty} \left| \left\langle \frac{w_{\text{prod}}}{\prod_{k=1}^L \|W_k\|_F}, v_1 \right\rangle \right| = 1,$$
and thus $\lim_{t\to\infty} \|w_{\text{prod}}\|_2 = \infty$.

Theorem 5.1 is proved using two lemmas, which may be of independent interest. To show the ideas, let us first introduce a little more notation. Recall that $\widehat{\mathcal{R}}(W)$ denotes the empirical risk induced by the deep linear network W. Abusing the notation a little, for any linear predictor $w \in \mathbb{R}^d$, we also use $\widehat{\mathcal{R}}(w)$ to denote the risk induced by w. With this notation, $\widehat{\mathcal{R}}(W) = \widehat{\mathcal{R}}(w_{\text{prod}})$, while

$$\nabla \widehat{\mathcal{R}}(w_{\text{prod}}) = \frac{1}{n} \sum_{i=1}^n \ell' \left(\langle w_{\text{prod}}, z_i \rangle \right) z_i = \frac{1}{n} \sum_{i=1}^n \ell' \left(W_L \cdots W_1 z_i \right) z_i$$

is in $\mathbb{R}^d$ and different from $\nabla \widehat{\mathcal{R}}(W)$, which has $\sum_{k=1}^L d_k d_{k-1}$ entries, as given below:

$$\frac{\partial \widehat{\mathcal{R}}}{\partial W_k} = W_{k+1}^\top \cdots W_L^\top \nabla \widehat{\mathcal{R}}(w_{\text{prod}})^\top W_1^\top \cdots W_{k-1}^\top.$$

Furthermore, for any $R > 0$, let

$$B(R) = \left\{ W \middle| \max_{1\le k\le L} \|W_k\|_F \le R \right\}.$$

First we show that for any $R > 0$, the time spent by gradient flow in $B(R)$ is finite.

Lemma 5.1. Under Assumptions 5.1 and 5.2, for any $R > 0$, there exists a constant $\epsilon(R) > 0$, such that for any $t \ge 1$ and any $W \in B(R)$, $\|\partial \widehat{\mathcal{R}} / \partial W_1\|_F \ge \epsilon(R)$. As a result, gradient flow spends a finite amount of time in $B(R)$ for any $R > 0$, and $\max_{1\le k\le L} \|W_k\|_F$ is unbounded.

Proof of Lemma 5.1. Fix an arbitrary $R > 0$. If the claim is not true, then for any $\epsilon > 0$, there exists some $t \geq 1$ such that $\|W_k\|_F \leq R$ for all k while $\left\|\partial\widehat{\mathcal{R}}/\partial W_1\right\|_F^2 \leq \epsilon^2$, which means

$$\left\|\frac{\partial\widehat{\mathcal{R}}}{\partial W_1}\right\|_F^2 = \left\|W_2^\top \cdots W_L^\top \nabla\widehat{\mathcal{R}}(w_{\text{prod}})^\top\right\|_F^2 = \|W_L \cdots W_2\|_2^2 \left\|\nabla\widehat{\mathcal{R}}(w_{\text{prod}})\right\|_2^2 \leq \epsilon^2.$$

Since $\|w_{\text{prod}}\|_2 \leq R^L$, we have

$$\langle \nabla\widehat{\mathcal{R}}(w_{\text{prod}}), u^*\rangle = \frac{1}{n}\sum_{i=1}^n \ell'\left(\langle w_{\text{prod}}, z_i\rangle\right)\langle z_i, u^*\rangle \leq \frac{1}{n}\sum_{i=1}^n \ell'\left(\langle w_{\text{prod}}, z_i\rangle\right)\gamma^* \leq -M\gamma^*,$$

where $-M = \max_{-R^L \leq x \leq R^L} \ell'(x)$. Since ℓ' is continuous and the domain is bounded, the maximum is attained and negative, and thus $M > 0$. Therefore $\left\|\nabla\widehat{\mathcal{R}}(w_{\text{prod}})\right\|_2 \geq M\gamma^*$, and thus $\|W_L \cdots W_2\|_2 \leq \epsilon/M\gamma^*$. Since $\|W_1\|_F \leq R$, we further have $\|w_{\text{prod}}\|_2 \leq \epsilon R/M\gamma^*$. In other words, after $t = 1$, $\left\|w_{\text{prod}}\right\|_2$ may be arbitrarily small, which implies $\widehat{\mathcal{R}}\left(w_{\text{prod}}\right)$ can be arbitrarily close to $\widehat{\mathcal{R}}\left(0\right)$.

On the other hand, by Assumption 5.2, $\mathrm{d}\widehat{\mathcal{R}}(W)/\,\mathrm{d}t = -\|\nabla\widehat{\mathcal{R}}(W)\|_F^2 < 0$ at $t = 0$. This implies that $\widehat{\mathcal{R}}\left(W(1)\right) < \widehat{\mathcal{R}}\left(W(0)\right)$, and for any $t \geq 1$, $\widehat{\mathcal{R}}\left(W(t)\right) \leq \widehat{\mathcal{R}}\left(W(1)\right) < \widehat{\mathcal{R}}\left(W(0)\right) \leq \widehat{\mathcal{R}}(0)$, which is a contradiction.

Since the risk is always positive, we have

$$\begin{aligned}
\widehat{\mathcal{R}}\left(W(0)\right) &\geq \int_{t=0}^\infty \sum_{k=1}^L \left\|\frac{\partial\widehat{\mathcal{R}}}{\partial W_k}\right\|_F^2 \mathrm{d}t \\
&\geq \int_{t=0}^\infty \left\|\frac{\partial\widehat{\mathcal{R}}}{\partial W_1}\right\|_F^2 \mathrm{d}t \\
&\geq \int_{t=0}^\infty \left\|\frac{\partial\widehat{\mathcal{R}}}{\partial W_1}\right\|_F^2 \mathbb{1}\left[\max_{1\leq k\leq L}\|W_k\|_F \leq R\right] \mathrm{d}t \\
&\geq \int_{t=1}^\infty \left\|\frac{\partial\widehat{\mathcal{R}}}{\partial W_1}\right\|_F^2 \mathbb{1}\left[\max_{1\leq k\leq L}\|W_k\|_F \leq R\right] \mathrm{d}t \\
&\geq \epsilon(R)^2 \int_{t=1}^\infty \mathbb{1}\left[\max_{1\leq k\leq L}\|W_k\|_F \leq R\right] \mathrm{d}t,
\end{aligned}$$

which implies gradient flow spends a finite amount of time in $\left\{W \middle| \max_{1\leq k\leq L}\|W_k\|_F \leq R\right\}$. This directly implies that $\max_{1\leq k\leq L}\|W_k\|_F$ is unbounded. QED.

To proceed, we need the following properties of linear networks from prior work [98, 102].

For any time $t \geq 0$ and any $1 \leq k < L$,

$$W_{k+1}^\top(t)W_{k+1}(t) - W_{k+1}^\top(0)W_{k+1}(0) = W_k(t)W_k^\top(t) - W_k(0)W_k^\top(0). \tag{5.2}$$

To see this, just notice that

$$W_{k+1}^\top \frac{\partial \widehat{\mathcal{R}}}{\partial W_{k+1}} = W_{k+1}^\top \cdots W_L^\top \nabla \widehat{\mathcal{R}}(w_{\text{prod}})^\top W_1^\top \cdots W_k^\top = \frac{\partial \widehat{\mathcal{R}}}{\partial W_k} W_k^\top.$$

Taking the trace on both sides of eq. (5.2), we have

$$\left\|W_{k+1}(t)\right\|_F^2 - \left\|W_{k+1}(0)\right\|_F^2 = \left\|W_k(t)\right\|_F^2 - \left\|W_k(0)\right\|_F^2. \tag{5.3}$$

In other words, the difference between the squares of Frobenius norms of any two layers remains a constant. Together with Lemma 5.1, it implies that all $\|W_k\|_F$ are unbounded.

However, even if $\|W_k\|_F$ are large, it does not follow necessarily that $\|w_{\text{prod}}\|_2$ is also large. Lemma 5.2 shows that this is indeed true: for gradient flow, as $\|W_k\|_F$ get larger, adjacent layers also get more aligned to each other, which ensures that their product also has a large norm.

Given a matrix W, let $\|W\|_\sigma$ denote its top singular value (the spectral norm). For $1 \leq k \leq L$, let σ_k, u_k, and v_k denote the first singular value, the first left singular vector, and the first right singular vector of W_k, respectively. Furthermore, define

$$D := \left(\max_{1\leq k\leq L} \|W_k(0)\|_F^2\right) - \|W_L(0)\|_F^2 + \sum_{k=1}^{L-1} \left\|W_k(0)W_k^\top(0) - W_{k+1}^\top(0)W_{k+1}(0)\right\|_\sigma,$$

which depends only on the initialization. If for any $1 \leq k < L$, it actually holds that $W_k(0)W_k^\top(0) = W_{k+1}^\top(0)W_{k+1}(0)$, then $D = 0$.

Lemma 5.2. The gradient flow iterates satisfy the following properties:

- For any $1 \leq k \leq L$, $\|W_k\|_F^2 - \|W_k\|_\sigma^2 \leq D$.
- For any $1 \leq k < L$, $\langle v_{k+1}, u_k\rangle^2 \geq 1 - \left(D + \|W_{k+1}(0)\|_\sigma^2 + \|W_k(0)\|_\sigma^2\right)/\|W_{k+1}\|_\sigma^2$.
- Suppose $\max_{1\leq k\leq L} \|W_k\|_F \to \infty$, then $\left|\left\langle w_{\text{prod}}/\prod_{k=1}^L \|W_k\|_F, v_1\right\rangle\right| \to 1$.

Proof. The first claim is true for $k = L$ since W_L is a row vector. For any $1 \leq k < L$, recall that [98, 102] give the following relation:

$$W_{k+1}^\top(t)W_{k+1}(t) - W_{k+1}^\top(0)W_{k+1}(0) = W_k(t)W_k^\top(t) - W_k(0)W_k^\top(0). \tag{5.4}$$

Let $A_{k,k+1} = W_k(0)W_k^\top(0) - W_{k+1}^\top(0)W_{k+1}(0)$. By eq. (5.4) and the definition of singular vectors and singular values, we have

$$\begin{aligned}
\sigma_k^2 &\geq v_{k+1}^\top W_k W_k^\top v_{k+1} \\
&= v_{k+1}^\top W_{k+1}^\top W_{k+1} v_{k+1} + v_{k+1}^\top A_{k,k+1} v_{k+1} \\
&= \sigma_{k+1}^2 + v_{k+1}^\top A_{k,k+1} v_{k+1} \\
&\geq \sigma_{k+1}^2 - \|A_{k,k+1}\|_\sigma.
\end{aligned} \tag{5.5}$$

Moreover, by taking the trace on both sides of eq. (5.4), we have

$$\begin{aligned}
\|W_k\|_F^2 = \operatorname{tr}\left(W_k W_k^\top\right) &= \operatorname{tr}\left(W_{k+1}^\top W_{k+1}\right) + \operatorname{tr}\left(W_k(0)W_k^\top(0)\right) - \operatorname{tr}\left(W_{k+1}^\top(0)W_{k+1}(0)\right) \\
&= \|W_{k+1}\|_F^2 + \|W_k(0)\|_F^2 - \|W_{k+1}(0)\|_F^2.
\end{aligned} \tag{5.6}$$

Summing eq. (5.5) and eq. (5.6) from k to $L-1$, we get

$$\|W_k\|_F^2 - \|W_k\|_\sigma^2 \leq \|W_k(0)\|_F^2 - \|W_L(0)\|_F^2 + \sum_{k'=k}^{L-1} \|A_{k',k'+1}\|_\sigma \leq D. \tag{5.7}$$

Next we prove that singular vectors get aligned. Consider $u_k^\top W_{k+1}^\top W_{k+1} u_k$. On one hand, similarly to eq. (5.5), we can get that

$$\begin{aligned}
u_k^\top W_{k+1}^\top W_{k+1} u_k &= u_k^\top W_k W_k^\top u_k - u_k^\top W_k(0)W_k^\top(0)u_k + u_k^\top W_{k+1}^\top(0)W_{k+1}(0)u_k \\
&\geq u_k^\top W_k W_k^\top u_k - u_k^\top W_k(0)W_k^\top(0)u_k \\
&\geq \sigma_k^2 - \|W_k(0)\|_\sigma^2.
\end{aligned} \tag{5.8}$$

On the other hand, it follows from the definition of singular vectors and eq. (5.7) that

$$\begin{aligned}
u_k^\top W_{k+1}^\top W_{k+1} u_k &= \langle u_k, v_{k+1}\rangle^2 \sigma_{k+1}^2 + u_k^\top \left(W_{k+1}^\top W_{k+1} - v_{k+1}\sigma_{k+1}^2 v_{k+1}^\top\right) u_k \\
&\leq \langle u_k, v_{k+1}\rangle^2 \sigma_{k+1}^2 + \|W_{k+1}\|_F^2 - \|W_{k+1}\|_\sigma^2 \\
&\leq \langle u_k, v_{k+1}\rangle^2 \sigma_{k+1}^2 + D.
\end{aligned} \tag{5.9}$$

Combining eq. (5.8) and eq. (5.9), we get

$$\sigma_k^2 \leq \langle u_k, v_{k+1}\rangle^2 \sigma_{k+1}^2 + D + \|W_k(0)\|_\sigma^2. \tag{5.10}$$

Similarly to eq. (5.8), we can get $\sigma_k^2 \geq v_{k+1}^\top W_k W_k^\top v_{k+1} \geq \sigma_{k+1}^2 - \|W_{k+1}(0)\|_\sigma^2$, which further

implies

$$\frac{\sigma_k^2}{\sigma_{k+1}^2} \geq 1 - \frac{\|W_{k+1}(0)\|_\sigma^2}{\sigma_{k+1}^2}. \tag{5.11}$$

Combining eq. (5.10) and eq. (5.11), we finally get

$$\langle u_k, v_{k+1} \rangle^2 \geq 1 - \frac{D + \|W_k(0)\|_\sigma^2 + \|W_{k+1}(0)\|_\sigma^2}{\sigma_{k+1}^2}.$$

Regarding the last claim, first recall that since the difference between the squares of Frobenius norms of any two layers is a constant, $\max_{1 \leq k \leq L} \|W_k\|_F \to \infty$ implies $\|W_k\|_F \to \infty$ for any k. We further have the following.

- Since $\|W_k\|_F^2 - \|W_k\|_\sigma^2 \leq D$, $\|W_k\|_2 \to \infty$ for any k, and $W_k / \|W_k\|_F \to u_k v_k^\top$.
- Since $\|W_k\|_\sigma \to \infty$, $|\langle u_k, v_{k+1} \rangle| \to 1$.

As a result,

$$\begin{aligned}
\left| \left\langle \frac{w_{\text{prod}}}{\prod_{k=1}^L \|W_k\|_F}, v_1 \right\rangle \right| &= \left| \left\langle \prod_{k=1}^L \frac{W_k}{\|W_k\|_F}, v_1 \right\rangle \right| \\
&\to \left| \left\langle \prod_{k=1}^L u_i v_i^\top, v_1 \right\rangle \right| \\
&\to 1.
\end{aligned}$$

QED.

Now we are ready to prove Theorem 5.1.

Proof. Suppose for some $\epsilon > 0$, $\widehat{\mathcal{R}}(W) \geq \epsilon$ for any t. Then there exists some $1 \leq j \leq n$ such that $\ell\left(\langle w_{\text{prod}}, z_j \rangle\right) \geq \epsilon$, and thus $\langle w_{\text{prod}}, z_j \rangle \leq \ell^{-1}(\epsilon)$. On the other hand, since $\widehat{\mathcal{R}}(W) \leq \widehat{\mathcal{R}}(0) = \ell(0)$, $\ell\left(\langle w_{\text{prod}}, z_j \rangle\right) \leq n\ell(0)$, and thus $\langle w_{\text{prod}}, z_j \rangle \geq \ell^{-1}\left(n\ell(0)\right)$. Let $-M = \max_{\ell^{-1}\left(n\ell(0)\right) \leq x \leq \ell^{-1}(\epsilon/n)} \ell'(x) < 0$, we have for any t,

$$\begin{aligned}
\langle \nabla \widehat{\mathcal{R}}(w_{\text{prod}}), u^* \rangle = \frac{1}{n} \sum_{i=1}^n \ell'\left(\langle w_{\text{prod}}, z_i \rangle\right) \langle z_i, u^* \rangle &\leq \frac{1}{n} \sum_{i=1}^n \ell'\left(\langle w_{\text{prod}}, z_i \rangle\right) \gamma^* \\
&\leq \frac{1}{n} \ell'\left(\langle w_{\text{prod}}, z_j \rangle\right) \gamma^* \\
&\leq \frac{-M\gamma^*}{n} < 0,
\end{aligned}$$

and thus $\|\nabla\widehat{\mathcal{R}}(w_{\text{prod}})\|_2 \geq M\gamma^*/n$.

Similar to the proof of Lemma 5.2, we can show that if $\|W_k\|_F \to \infty$,

$$\left|\left\langle \frac{(W_L \cdots W_2)^\top}{\|W_k\|_F \cdots \|W_2\|_F}, v_2 \right\rangle\right| \to 1.$$

In other words, there exists some $C > 0$, such that when $\min_{1\leq k\leq L} \|W_k\|_F > C$, it holds that $\|W_L \cdots W_2\|_2 \geq \|W_k\|_F \cdots \|W_2\|_F/2 > C^L/2$.

Lemma 5.1 shows that gradient flow spends finite time in $\left\{W \middle| \max_{1\leq k\leq L} \|W_k\|_F \leq R\right\}$ for any $R > 0$. Since the difference between the squares of Frobenius norms of any two layers is a constant, gradient flow also spends a finite amount of time in $\left\{W \middle| \min_{1\leq k\leq L} \|W_k\|_F \leq C\right\}$. Now we have

$$\begin{aligned}
\widehat{\mathcal{R}}\left(W(0)\right) &\geq \int_{t=0}^{\infty} \sum_{k=1}^{L} \left\| \frac{\partial \widehat{\mathcal{R}}}{\partial W_k} \right\|_F^2 \mathrm{d}t \\
&\geq \int_{t=0}^{\infty} \left\| \frac{\partial \widehat{\mathcal{R}}}{\partial W_1} \right\|_F^2 \mathrm{d}t \\
&= \int_{t=0}^{\infty} \|W_L \cdots W_2\|_2^2 \|\nabla\widehat{\mathcal{R}}(w_{\text{prod}})\|_2^2 \,\mathrm{d}t \\
&\geq \int_{t=0}^{\infty} \|W_L \cdots W_2\|_2^2 \|\nabla\widehat{\mathcal{R}}(w_{\text{prod}})\|_2^2 \mathbb{1}\left[W \middle| \min_{1\leq k\leq L} \|W_k\|_F > C\right] \mathrm{d}t \\
&\geq \left(\frac{M\gamma^*}{n}\right)^2 \left(\frac{C^L}{2}\right)^2 \int_{t=0}^{\infty} \mathbb{1}\left[W \middle| \min_{1\leq k\leq L} \|W_k\|_F > C\right] \mathrm{d}t \\
&= \infty,
\end{aligned}$$

which is a contradiction. Therefore $\widehat{\mathcal{R}}(\epsilon) \to 0$. This further implies $\|W_k\|_F \to \infty$, since $\widehat{\mathcal{R}}(W)$ has no finite optimum. Finally, invoking Lemma 5.2 proves the final claim of Theorem 5.1. QED.

5.1.2 Margin maximization

In [31], we also proved the following result.

Theorem 5.2. Suppose Assumption 5.2 and that the support vectors span $\mathbb{R}^d$, for almost all data and the exponential loss or logistic loss, we have $\lim_{t\to\infty} \left|\langle v_1, \bar{u}\rangle\right| = 1$, where v_1 is the first right singular vector of W_1. As a result, $\lim_{t\to\infty} w_{\text{prod}}/\prod_{k=1}^{L} \|W_k\|_F = \bar{u}$.

Note that Theorem 5.2 requires that the support vectors span $\mathbb{R}^d$; later in Section 5.2.2, we will show the same result without this condition, as a corollary of a general alignment result for deep homogeneous networks.

5.2 ALIGNMENT IN DEEP HOMOGENEOUS NETWORKS

In this section, we consider L-homogeneous networks, meaning that given any input x and any positive number $c > 0$, it holds that $f(x; cW) = c^L f(x; W)$. Examples include deep networks, with linear and convolutional layers, max and average pooling layers, and homogeneous activations, such as the identity activation $x \mapsto x$, ReLU activation $x \mapsto \max\{0, x\}$, and more generally powers of ReLU $x \mapsto \max\{0, x\}^k$. On the other hand, homogeneity does not allow skip connections and bias vectors. Here is a typical homogeneous network, where the activation σ is homogeneous:

$$x \mapsto W_L \sigma \left(W_{L-1} \sigma \left(\cdots \sigma(W_1 x) \cdots \right) \right).$$

For simplicity, in the following we assume f is twice continuously differentiable, even though we can drop this condition by assuming certain definability conditions [32].

For simplicity, we will focus on the exponential loss, and let $\mathcal{L}$ denote the unnormalized empirical risk

$$\mathcal{L}(W) := \sum_{i=1}^{n} \ell \left(y_i f(x_i; W) \right) = \sum_{i=1}^{n} \exp \left(-y_i f(x_i; W) \right) = \sum_{i=1}^{n} \exp \left(-h_i(W) \right),$$

where we let $h_i(W) := y_i f(x_i; W)$. In the following, we will not deal with y_i and x_i directly, instead we will just consider h_i, which are also positive homogeneous and twice differentiable.

We consider gradient flow over $\mathcal{L}$:

$$\frac{\mathrm{d}W_t}{\mathrm{d}t} = -\nabla \mathcal{L}(W_t).$$

Note that gradient flows over $\widehat{\mathcal{R}}$ and $\mathcal{L}$ have the same path, and thus will not affect the results we prove below. We further make the following assumption on the initialization:

Assumption 5.3. The initial iterate W_0 satisfies $\mathcal{L}(W_0) < \ell(0)$.

Note that Assumption 5.3 is stronger than Assumption 5.2. However, we can ensure Assumption 5.3 using Theorem 5.1 or an NTK analysis.

In [32], we show the following result.

Theorem 5.3. Under Assumption 5.3, if the network is twice differentiable, then $-\nabla\mathcal{L}(W_t)$ and W_t become aligned to each other, meaning the angle between W_t and $-\nabla\mathcal{L}(W_t)$ converges to zero.

Previously, [103] showed that *subsequences* of the gradient flow converge to KKT points of the margin maximization problem. We note that alignment in Theorem 5.3 is in general a stronger notion, in that it is unclear how to prove alignment as a consequence of convergence to KKT points.

Below we first prove Theorem 5.3, and then apply it to show margin maximization for deep linear networks. For simplicity, in this section we let $\|\cdot\|$ denote the ℓ_2 norm of vectors or Frobenius norm of matrices.

5.2.1 Proof of Theorem 5.3

We first give the following technical result.

Lemma 5.3. Suppose $f : \mathbb{R}^k \to \mathbb{R}$ is differentiable and L-positively homogeneous for some $L > 0$. Then ∇f is $(L-1)$-positively homogeneous: given any nonzero x and $c > 0$, we have

$$\nabla f(cx) = c^{L-1}\nabla f(x).$$

If ∇f is also differentiable, then for any $c > 0$, it holds that

$$\nabla^2 f(cx) = c^{L-2}\nabla^2 f(x).$$

Moreover, there exists $K_\sigma > 0$ such that for any $\|x\| = 1$, it holds that $\left\|\nabla^2 f(x)\right\|_\sigma \le K_\sigma$.

Proof. By definition,

$$\lim_{\|y\|\downarrow 0} \frac{f(x+y) - f(x) - \left\langle \nabla f(x), y\right\rangle}{\|y\|} = 0.$$

On the other hand, by homogeneity,

$$f(cx+z) - f(cx) - \left\langle c^{L-1}\nabla f(x), z\right\rangle = c^L\left(f\left(x + \frac{z}{c}\right) - f(x) - \left\langle \nabla f(x), \frac{z}{c}\right\rangle\right).$$

Therefore

$$\lim_{\|z\|\downarrow 0} \frac{f(cx+z) - f(cx) - \left\langle c^{L-1}\nabla f(x), z\right\rangle}{\|z\|} = c^{L-1}\lim_{\|z\|\downarrow 0} \frac{f\left(x+\frac{z}{c}\right) - f(x) - \left\langle \nabla f(x), \frac{z}{c}\right\rangle}{\|z/c\|} = 0,$$

which proves the claim. The homogeneity of $\nabla^2 f$ when it exists can be proved in the same way.

To get K_σ, note that for any $\|x\| = 1$, there exists an open neighborhood U_x of x on which ∇f is K_x-Lipschitz continuous, and thus the spectral norm of $\nabla^2 f$ is bounded by K_x. All the U_x form an open cover of the compact unit sphere, and thus has a finite subcover, which implies the claim. QED.

We also define the following quantities, which will be useful in our analysis. Let

$$\alpha(W) := \ell^{-1}\big(\mathcal{L}(W)\big) = -\ln\big(\mathcal{L}(W)\big), \quad \text{and} \quad \beta(W) := \frac{\big\langle W, \nabla\alpha(W)\big\rangle}{L}.$$

Lemma 5.4. If $\mathcal{L}(W) < \ell(0)$, it holds that

$$0 < \alpha(W) \le \min_{1\le i\le n} h_i(W) \le \beta(W) \le \alpha(W) + \ln(n).$$

Proof. Note that $\alpha(W) = -\ln\big(\mathcal{L}(W)\big) > -\ln\big(\ell(0)\big) = 0$. Moreover,

$$\beta(W) = \frac{1}{L}\sum_{i=1}^n \frac{\exp\big(-h_i(W)\big)}{\sum_{i'=1}^n \exp\big(-h_{i'}(W)\big)} \big\langle W, \nabla h_i(W)\big\rangle = \sum_{i=1}^n \frac{\exp\big(-h_i(W)\big)}{\sum_{i'=1}^n \exp\big(-h_{i'}(W)\big)} h_i(W),$$

where we use Euler's homogeneous function theorem. Therefore

$$\alpha(W) = -\ln\left(\sum_{i=1}^n \exp\big(-h_i(W)\big)\right) \le \min_{1\le i\le n} h_i(W) \le \beta(W).$$

Finally, since the ln-sum-exp function is convex, we have

$$\begin{aligned}\ln(n) + \alpha(W) &= \ln\left(\sum_{i=1}^n \exp(0)\right) - \ln\left(\sum_{i=1}^n \exp\big(-h_i(W)\big)\right) \\ &\ge \sum_{i=1}^n \frac{\exp\big(-h_i(W)\big)}{\sum_{i'=1}^n \exp\big(-h_{i'}(W)\big)} h_i(W) = \beta(W).\end{aligned}$$

QED.

Next we estimate various quantities using Lemmas 5.3 and 5.4.

Lemma 5.5. For any W which satisfies $\mathcal{L}(W) < \ell(0)$, it holds that $\beta(W)/\|W\|^L$ and $\big\|\nabla\alpha(W)\big\|/\|W\|^{L-1}$ are bounded.

Proof. Since $h_i(W)$ is continuous, it is bounded on the unit sphere. Because it is L-positively homogeneous, $h_i(W)/\|W\|^L$ is bounded on $\mathbb{R}^k$. Lemma 5.4 implies that $\beta(W) - \ln(n) \leq \alpha(W) \leq \min_{1\leq i\leq n} h_i(W)$, and it follows that $\beta(W)/\|W\|^L$ is bounded.

Recall that

$$\nabla\alpha(W) = \sum_{i=1}^{n} \frac{\exp\big(-h_i(W)\big)}{\sum_{i'=1}^{n}\exp\big(-h_{i'}(W)\big)}\nabla h_i(W),$$

Moreover, Lemma 5.3 implies that all $\big\|\nabla h_i(W)\big\|/\|W\|^{L-1}$ are bounded. Consequently, $\big\|\nabla\alpha(W)\big\|/\|W\|^{L-1}$ is bounded. QED.

Next we define the following quantity $\mathcal{J}$; it generalizes the dual objective in the linear case (cf. Section 2.2), and is crucial in our analysis.

$$\mathcal{J}(W) := \frac{\big\|\nabla\alpha(W)\big\|^2}{\|W\|^{2L-2}}. \tag{5.12}$$

Lemma 5.6. For any W satisfying $\mathcal{L}(W) < \ell(0)$,

$$\big\langle \nabla\mathcal{J}(W), -\nabla\mathcal{L}(W)\big\rangle \leq K\mathcal{L}(W)\|W\|^{L-2}\sin(\theta)^2$$

for some constant $K > 0$, where θ denotes the angle between W and $-\nabla\mathcal{L}(W)$.

Proof. For simplicity, let $\pi(\xi) := -\ln\big(\sum_{i=1}^{n}\exp(-\xi_i)\big)$. It holds that π is concave, and

$$\nabla\alpha(W) = \sum_{i=1}^{n}\frac{\partial\pi}{\partial h_i}\nabla h_i(W). \tag{5.13}$$

Since ∇h_i are also differentiable at W, we have

$$\nabla^2\alpha(W) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial^2\pi}{\partial h_i\partial h_j}\nabla h_i(W)\nabla h_j(W)^\top\right) + \sum_{i=1}^{n}\frac{\partial\pi}{\partial h_i}\nabla^2 h_i(W). \tag{5.14}$$

On the other hand, let $\widetilde{W} = W/\|W\|$, we have

$$\begin{aligned}\nabla\mathcal{J}(W) &= \frac{2\nabla^2\alpha(W)\nabla\alpha(W)}{\|W\|^{2L-2}} - \frac{\big\|\nabla\alpha(W)\big\|^2}{\|W\|^{4L-4}}\cdot(2L-2)\|W\|^{2L-3}\widetilde{W}\\ &= \frac{2\nabla^2\alpha(W)\nabla\alpha(W)}{\|W\|^{2L-2}} - \frac{(2L-2)\big\|\nabla\alpha(W)\big\|^2}{\|W\|^{2L}}W,\end{aligned}$$

and thus

$$\begin{aligned}
&\frac{\|W\|^{2L}}{2}\frac{\left\langle \nabla\mathcal{J}(W), -\nabla\mathcal{L}(W)\right\rangle}{\mathcal{L}(W)} \\
&= \frac{\|W\|^{2L}}{2}\left\langle \nabla\mathcal{J}(W), \nabla\alpha(W)\right\rangle \\
&= \|W\|^2\nabla\alpha(W)^\top\nabla^2\alpha(W)\nabla\alpha(W) - (L-1)\left\|\nabla\alpha(W)\right\|^2\left\langle W, \nabla\alpha(W)\right\rangle. \qquad (5.15)
\end{aligned}$$

Comparing eqs. (5.14) and (5.15), first note that

$$\sum_{i=1}^n\sum_{j=1}^n \frac{\partial^2\pi}{\partial h_i\partial h_j}\nabla\alpha(W)^\top\nabla h_i(W)\nabla h_j(W)^\top\nabla\alpha(W) \le 0,$$

since π is concave. Moreover by eq. (5.13),

$$\left\langle W, \nabla\alpha(W)\right\rangle = \sum_{i=1}^n \frac{\partial\pi}{\partial h_i}\left\langle W, \nabla h_i(W)\right\rangle = L\sum_{i=1}^n \frac{\partial\pi}{\partial h_i}h_i(W).$$

Therefore eq. (5.15) is upper bounded by

$$\|W\|^2\sum_{i=1}^n \frac{\partial\pi}{\partial h_i}\nabla\alpha(W)^\top\nabla^2 h_i(W)\nabla\alpha(W) - L(L-1)\left\|\nabla\alpha(W)\right\|^2\sum_{i=1}^n \frac{\partial\pi}{\partial h_i}h_i(W). \qquad (5.16)$$

Let $\nabla_r\alpha(W)$ and $\nabla_\perp\alpha(W)$ denote the radial and spherical part of $\nabla\alpha(W)$. Let θ denote the angle between W and $\nabla\alpha(W)$. Lemma 5.4 and the definition of $\beta(W)$ imply that

$$\left\langle W, \nabla\alpha(W)\right\rangle = L\beta(W) > 0,$$

and thus θ is between 0 and $\pi/2$. Now Lemma 5.3 implies that

$$\begin{aligned}
\|W\|^2\nabla_r\alpha(W)^\top\nabla^2 h_i(W)\nabla_r\alpha(W) &= \cos(\theta)^2\left\|\nabla\alpha(W)\right\|^2 W^\top\nabla^2 h_i(W)W \\
&= \cos(\theta)^2\left\|\nabla\alpha(W)\right\|^2\cdot L(L-1)h_i(W) \\
&\le \left\|\nabla\alpha(W)\right\|^2\cdot L(L-1)h_i(W). \qquad (5.17)
\end{aligned}$$

Moreover,

$$\begin{aligned}
2\|W\|^2\nabla_\perp\alpha(W)^\top\nabla^2 h_i(W)\nabla_r\alpha(W) &= 2\|W\|\left\|\nabla\alpha(W)\right\|\cos(\theta)\left\langle \nabla_\perp\alpha(W), \nabla^2 h_i(W)W\right\rangle \\
&= 2(L-1)\|W\|\left\|\nabla\alpha(W)\right\|\cos(\theta)\left\langle \nabla_\perp\alpha(W), \nabla h_i(W)\right\rangle,
\end{aligned}$$

and thus by the definition of $\beta(W)$,

$$\begin{aligned}
&2\|W\|^2 \sum_{i=1}^{n} \frac{\partial \pi}{\partial h_i} \nabla_\perp \alpha(W)^\top \nabla^2 h_i(W) \nabla_r \alpha(W) \\
&= 2(L-1)\|W\| \big\|\nabla \alpha(W)\big\| \cos(\theta) \left\langle \nabla_\perp \alpha(W), \nabla \alpha(W) \right\rangle \\
&= 2(L-1)\|W\| \big\|\nabla \alpha(W)\big\|^3 \cos(\theta) \sin(\theta)^2 \\
&= 2L(L-1)\big\|\nabla \alpha(W)\big\|^2 \sin(\theta)^2 \beta(W). \qquad (5.18)
\end{aligned}$$

In addition, Lemma 5.3 ensures that $\|\nabla^2 f\|_\sigma$ has a uniform bound K_σ on the unit sphere, therefore

$$\begin{aligned}
\|W\|^2 \sum_{i=1}^{n} \frac{\partial \pi}{\partial h_i} \nabla_\perp \alpha(W)^\top \nabla^2 h_i(W) \nabla_\perp \alpha(W) &\le \|W\|^2 \big\|\nabla \alpha(W)\big\|^2 \sin(\theta)^2 \cdot K_\sigma \|W\|^{L-2} \\
&= K_\sigma \|W\|^L \big\|\nabla \alpha(W)\big\|^2 \sin(\theta)^2. \qquad (5.19)
\end{aligned}$$

Combining eqs. (5.15) to (5.19) gives

$$\frac{\left\langle \nabla \mathcal{J}(W), -\nabla \mathcal{L}(W) \right\rangle}{\mathcal{L}(W)} \le \frac{4\left(K_\sigma \|W\|^L + L(L-1)\beta(W)\right)\big\|\nabla \alpha(W)\big\|^2}{\|W\|^{2L}} \sin(\theta)^2.$$

Invoking Lemma 5.5 then gives

$$\left\langle \nabla \mathcal{J}(W), -\nabla \mathcal{L}(W) \right\rangle \le -K \mathcal{L}(W) \|W\|^{L-2} \sin(\theta)^2$$

for some constant $K > 0$. QED.

To continue, we need a little more notation. Let $\tilde{\alpha}(W) := \frac{\alpha(W)}{\|W\|^L}$. Next we note that the gradients of α and $\tilde{\alpha}$ are strongly related.

Lemma 5.7. For any nonzero $W \in \mathbb{R}^k$, we have

$$\nabla_r \tilde{\alpha}(W) = L \frac{\beta(W) - \alpha(W)}{\|W\|^{L+1}} \widetilde{W}, \quad \text{and} \quad \nabla_\perp \tilde{\alpha}(W) = \frac{\nabla_\perp \alpha(W)}{\|W\|^L}.$$

Moreover,

$$\frac{\mathrm{d}\tilde{\alpha}_t}{\mathrm{d}t} = \big\|\nabla_r \tilde{\alpha}(W_t)\big\| \big\|\nabla_r \mathcal{L}(W_t)\big\| + \big\|\nabla_\perp \tilde{\alpha}(W_t)\big\| \big\|\nabla_\perp \mathcal{L}(W_t)\big\|.$$

Proof. Note that given $W \neq 0$, α is differentiable at W if and only if $\tilde{\alpha}$ is differentiable at

W, and when both gradients exist,

$$\nabla\tilde{\alpha}(W) = \frac{\nabla\alpha(W)}{\|W\|^L} - \frac{\alpha(W)\cdot L\|W\|^{L-1}\widetilde{W}}{\|W\|^{2L}} = \frac{\nabla\alpha(W)}{\|W\|^L} - L\frac{\alpha(W)\widetilde{W}}{\|W\|^{L+1}}.$$

The first claim of Lemma 5.7 then follows from the definition of $\beta(W)$. The second claim is trivial. For the final claim, note that

$$\frac{\mathrm{d}\tilde{\alpha}(W_t)}{\mathrm{d}t} = \left\langle \nabla\tilde{\alpha}(W_t), -\nabla\mathcal{L}(W_t)\right\rangle = \left\langle \nabla_r\tilde{\alpha}(W_t), -\nabla_r\mathcal{L}(W_t)\right\rangle + \left\langle \nabla_\perp\tilde{\alpha}(W_t), -\nabla_\perp\mathcal{L}(W_t)\right\rangle.$$

By Lemma 5.4 and the first claim of Lemma 5.7, both $\left\langle \nabla_r\tilde{\alpha}(W_t), \widetilde{W_t}\right\rangle$ and $\left\langle -\nabla_r\mathcal{L}(W_t), \widetilde{W_t}\right\rangle$ are nonnegative, and thus

$$\left\langle \nabla_r\tilde{\alpha}(W_t), -\nabla_r\mathcal{L}(W_t)\right\rangle = \left\|\nabla_r\tilde{\alpha}(W_t)\right\|\left\|\nabla_r\mathcal{L}(W_t)\right\|.$$

The second claim of Lemma 5.7 also implies that $\nabla_\perp\tilde{\alpha}(W_t)$ and $-\nabla_\perp\mathcal{L}(W_t)$ point to the same direction, and thus

$$\left\langle \nabla_\perp\tilde{\alpha}(W_t), -\nabla_\perp\mathcal{L}(W_t)\right\rangle = \left\|\nabla_\perp\tilde{\alpha}(W_t)\right\|\left\|\nabla_\perp\mathcal{L}(W_t)\right\|.$$

QED.

Next we control θ_t, the angle between W_t and $-\nabla\mathcal{L}(W_t)$, using Lemma 5.7.

Lemma 5.8. If $\mathcal{L}(W_0) < \ell(0)$, then it holds that

$$\int_0^\infty \mathcal{L}(W_t)\|W_t\|^{L-2}\tan(\theta_t)^2\,\mathrm{d}t < \infty.$$

Proof. Lemma 5.7 implies that

$$\frac{\mathrm{d}\tilde{\alpha}_t}{\mathrm{d}t} \geq \left\|\nabla_\perp\tilde{\alpha}(W_t)\right\|\left\|\nabla_\perp\mathcal{L}(W_t)\right\| = \frac{\left\|\nabla_\perp\alpha(W_t)\right\|\left\|\nabla_\perp\mathcal{L}(W_t)\right\|}{\|W_t\|^L} = \frac{\mathcal{L}(W_t)\left\|\nabla_\perp\alpha(W_t)\right\|^2}{\|W_t\|^L},$$

and moreover $\left\|\nabla_\perp\alpha(W_t)\right\| = \left\|\nabla_r\alpha(W_t)\right\|\tan(\theta_t) = \frac{L\beta(W_t)}{\|W_t\|}\tan(\theta_t)$, therefore

$$\frac{\mathrm{d}\tilde{\alpha}_t}{\mathrm{d}t} \geq \mathcal{L}(W_t)\cdot L^2\tan(\theta_t)^2\frac{\beta(W_t)^2}{\|W_t\|^{L+2}}.$$

Since $\tilde{\alpha}_t$ is monotonically nondecreasing with a limit a [103, Theorem 4.1], it also follows that $\beta(W_t)/\|W_t\|^L \geq \tilde{\alpha}(W_t) \geq \tilde{\alpha}(W_0) > 0$, and the proof is finished. QED.

Now we can prove Theorem 5.3.

Proof of Theorem 5.3. Fix an arbitrary $\epsilon \in (0,1)$, and let $\mathcal{J}_t$ denote $\mathcal{J}(W_t)$. Recall that $\lim_{t\to\infty} \alpha(W_t)/\|W_t\|^L = a$. Lemma 5.4 then implies $\lim_{t\to\infty} \beta(W_t)/\|W_t\|^L = a$, and thus we can find t_1 such that for any $t > t_1$,

$$a\left(1-\frac{\epsilon}{6}\right) < \frac{\beta(W_t)}{\|W_t\|^L} = \frac{1}{L}\left\langle \frac{\nabla\alpha(W_t)}{\|W_t\|^{L-1}}, \frac{W_t}{\|W_t\|_F}\right\rangle < a\left(1+\frac{\epsilon}{6}\right). \tag{5.20}$$

Moreover, Lemmas 5.6 and 5.8 imply that there exists t_2 such that for any $t' > t > t_2$,

$$\mathcal{J}_{t'} - \mathcal{J}_t < \left(\frac{aL\epsilon}{6}\right)^2. \tag{5.21}$$

[103, Corollary C.10] implies that there exists $t_3 > \max\{t_1, t_2\}$ such that

$$\frac{1}{\cos(\theta_{t_2})^2} - 1 < \frac{\epsilon}{3}, \quad \text{and thus} \quad \frac{1}{\cos(\theta_{t_2})} < 1 + \frac{\epsilon}{6}. \tag{5.22}$$

We claim that $\delta_t < 1+\epsilon$ for any $t > t_3$.

To see this, note that eqs. (5.20) and (5.22) imply

$$\sqrt{\mathcal{J}_{t_2}} = \frac{\left\|\nabla\alpha(W_{t_2})\right\|}{\|W_{t_2}\|^{L-1}} < aL\left(1+\frac{\epsilon}{6}\right)\frac{1}{\cos(\theta_{t_2})} < aL\left(1+\frac{\epsilon}{6}\right)^2 < aL\left(1+\frac{\epsilon}{2}\right).$$

Moreover, using eq. (5.21), for any $t > t_2$,

$$\sqrt{\mathcal{J}_t} = \sqrt{\mathcal{J}_{t_2} + \mathcal{J}_t - \mathcal{J}_{t_2}} < \sqrt{\mathcal{J}_{t_2} + \left(\frac{\gamma L\epsilon}{6}\right)^2} < \sqrt{\mathcal{J}_{t_2}} + \frac{aL\epsilon}{6} < aL\left(1+\frac{2\epsilon}{3}\right),$$

and thus

$$\frac{1}{\cos(\theta_t)} = \frac{\sqrt{\mathcal{J}_t}}{L\beta(W_t)/\|W_t\|^L} < \frac{aL\left(1+2\epsilon/3\right)}{aL(1-\epsilon/6)} < 1+\epsilon.$$

Since ϵ is arbitrary, we have $\lim_{t\to\infty}\theta_t = 0$. QED.

5.2.2 Application to deep linear networks

Recall that Theorem 5.1 implies w_{prod} becomes aligned with v_1, the top right singular vector of W_1. Moreover, since W_1 and $\partial\mathcal{L}/\partial W_1$ become aligned as ensured by Theorem 5.3, and since $\partial\widehat{\mathcal{R}}/\partial W_1 = W_2^\top \cdots W_L^\top \nabla\widehat{\mathcal{R}}(w_{\text{prod}})^\top$, it follows that v_1 becomes aligned with $\nabla\widehat{\mathcal{R}}(w_{\text{prod}})$,

and thus w_{prod} becomes aligned with $\nabla\widehat{\mathcal{R}}(w_{\text{prod}})$. In other words, w_{prod} asymptotically satisfies the duality condition of the margin maximization problem (cf. Lemma 2.3), and thus w_{prod} converges to the maximum-margin predictor.

5.2.3 Directional convergence

Note that Theorem 5.3 only shows W_t and $-\nabla\mathcal{L}(W_t)$ converges to the same direction, but it does not show if W_t itself converges to a fixed direction, i.e., if $W_t/\|W_t\|_F$ converges to a fixed point over the unit sphere. This property is called "directional convergence"; it is often assumed throughout the literature [96, 104], but has only been established for linear predictors [16]. It is tricky to prove because it may still be false for highly smooth functions: for instance, the homogeneous Mexican Hat function satisfies all our assumptions *except* definability, and can be adjusted to have arbitrary order of continuous derivatives, but its gradient flow *does not* converge in direction, instead it spirals [103]. In [32], we proved directional convergence under an additional assumption of o-minimal definability which is mild and satisfied by most practical neural networks; see [32, Appendix B] for details. As mentioned above, the twice differentiability condition in Theorem 5.3 can also be replaced with definability and locally Lipschitz gradients; see [32, Section 4] for details.

5.3 FUTURE DIRECTIONS

In this chapter, we summarize our implicit bias results of GD on deep networks. However, these results are still not satisfactory enough: even though they are true as $t \to \infty$, it may actually take a really long time for these alignment phenomena to happen. This is in contrast to the NTK-style analysis given in Chapter 4: an NTK analysis basically only uses the power of random features at initialization, but in practice these features do change during training, and they can often be better than the random features at initialization [105]. It is very interesting to formally study how the features evolve after the initial NTK phase of training, and understand when and why they are better than random features.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[3] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton et al., "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[4] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

[5] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[7] M. Anthony, P. L. Bartlett, P. L. Bartlett et al., *Neural network learning: Theoretical foundations*. cambridge university press Cambridge, 1999, vol. 9.

[8] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[9] S. Bubeck, "Convex optimization: Algorithms and complexity," *arXiv preprint arXiv:1405.4980*, 2014.

[10] A. B. Novikoff, "On convergence proofs for perceptrons," STANFORD RESEARCH INST MENLO PARK CA, Tech. Rep., 1963.

[11] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[12] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. MIT Press, 2012.

[13] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 6240–6249.

[14] Z. Ji and M. Telgarsky, "Risk and parameter convergence of logistic regression," *arXiv preprint arXiv:1803.07300v2*, 2018.

[15] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *ICML*. Citeseer, 2012.

[16] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[17] Z. Ji and M. Telgarsky, "Risk and parameter convergence of logistic regression," *arXiv preprint arXiv:1803.07300*, 2018.

[18] M. Telgarsky, "Margins, shrinkage, and boosting," in *ICML*, 2013.

[19] Z. Ji and M. Telgarsky, "Characterizing the implicit bias via a primal-dual analysis," *arXiv preprint arXiv:1906.04540*, 2019.

[20] Z. Ji, N. Srebro, and M. Telgarsky, "Fast margin maximization via dual acceleration," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4860–4869.

[21] A. Cotter, S. Shalev-Shwartz, and N. Srebro, "The kernelized stochastic batch perceptron," *arXiv preprint arXiv:1204.0566*, 2012.

[22] Z. Ji, M. Dudík, R. E. Schapire, and M. Telgarsky, "Gradient descent follows the regularization path for general losses," in *Conference on Learning Theory*. PMLR, 2020, pp. 2109–2136.

[23] P. Awasthi, M. F. Balcan, and P. M. Long, "The power of localization for efficiently learning linear separators with noise," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 2014, pp. 449–458.

[24] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis, "Non-convex sgd learns halfspaces with adversarial label noise," *arXiv preprint arXiv:2006.06742*, 2020.

[25] S. Frei, Y. Cao, and Q. Gu, "Agnostic learning of halfspaces with gradient descent via soft margins," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3417–3426.

[26] Z. Ji, K. Ahn, P. Awasthi, S. Kale, and S. Karp, "Agnostic learnability of halfspaces via logistic loss," *arXiv preprint arXiv:2201.13419*, 2022.

[27] Z. Ji, J. D. Li, and M. Telgarsky, "Early-stopped neural networks are consistent," *arXiv preprint arXiv:2106.05932*, 2021.

[28] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in neural information processing systems*, 2018, pp. 8571–8580.

[29] Z. Chen, Y. Cao, D. Zou, and Q. Gu, "How much over-parameterization is sufficient to learn deep relu networks?" *arXiv preprint arXiv:1911.12360*, 2019.

[30] Z. Ji and M. Telgarsky, "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks," *arXiv preprint arXiv:1909.12292*, 2019.

[31] Z. Ji and M. Telgarsky, "Gradient descent aligns the layers of deep linear networks," *arXiv preprint arXiv:1810.02032*, 2018.

[32] Z. Ji and M. Telgarsky, "Directional convergence and alignment in deep learning," *arXiv preprint arXiv:2006.06657*, 2020.

[33] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[34] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf, 2008.

[35] Z. Allen-Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," *arXiv preprint arXiv:1407.1537*, 2014.

[36] Y. Cao and Q. Gu, "Generalization error bounds of gradient descent for learning over-parameterized deep relu networks," *arXiv preprint arXiv:1902.01384*, 2019.

[37] O. Shamir, "Gradient methods never overfit on separable data," *arXiv preprint arXiv:2007.00028*, 2020.

[38] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Mathematical programming*, vol. 120, no. 1, pp. 221–259, 2009.

[39] J. Borwein and A. S. Lewis, *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[40] R. M. Freund, P. Grigas, and R. Mazumder, "Adaboost and forward stagewise regression are first-order convex optimization methods," *arXiv preprint arXiv:1307.1192*, 2013.

[41] S. Shalev-Shwartz and Y. Singer, "Online learning: Theory, algorithms, and applications," 2007.

[42] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[43] R. T. Rockafellar, *Convex analysis*. Princeton university press, 2015.

[44] M. S. Nacson, J. Lee, S. Gunasekar, N. Srebro, and D. Soudry, "Convergence of gradient descent on separable data," *arXiv preprint arXiv:1803.01905*, 2018.

[45] A. Ramdas and J. Pena, "Towards a deeper geometric, analytic and algorithmic understanding of margins," *Optimization Methods and Software*, vol. 31, no. 2, pp. 377–391, 2016.

[46] S. Rosset, J. Zhu, and T. Hastie, "Boosting as a regularized path to a maximum margin classifier," *JMLR*, vol. 5, pp. 941–973, 2004.

[47] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee et al., "Boosting the margin: A new explanation for the effectiveness of voting methods," *The annals of statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[48] T. Zhang and B. Yu, "Boosting with early stopping: Convergence and consistency," *The Annals of Statistics*, vol. 33, pp. 1538–1579, 2005.

[49] P. Zhao and B. Yu, "Stagewise lasso," *JMLR*, vol. 8, no. Dec, pp. 2701–2726, 2007.

[50] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing implicit bias in terms of optimization geometry," *arXiv preprint arXiv:1802.08246*, 2018.

[51] M. S. Nacson, N. Srebro, and D. Soudry, "Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 3051–3059.

[52] A. Suggala, A. Prasad, and P. K. Ravikumar, "Connecting optimization and regularization paths," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[53] Z. Charles, S. Rajput, S. Wright, and D. Papailiopoulos, "Convergence and margin of adversarial training on separable data," *arXiv preprint arXiv:1905.09209*, 2019.

[54] Y. Li, E. X. Fang, H. Xu, and T. Zhao, "Implicit bias of gradient descent based adversarial training on separable data," 2020.

[55] B. Wang, Q. Meng, H. Zhang, R. Sun, W. Chen, and Z.-M. Ma, "Momentum doesn't change the implicit bias," *arXiv preprint arXiv:2110.03891*, 2021.

[56] S. Ben-David, D. Loker, N. Srebro, and K. Sridharan, "Minimizing the misclassification error rate using a surrogate convex loss," *arXiv preprint arXiv:1206.6442*, 2012.

[57] A. Daniely, "A ptas for agnostically learning halfspaces," in *Conference on Learning Theory*. PMLR, 2015, pp. 484–502.

[58] M. J. Kearns, R. E. Schapire, and L. M. Sellie, "Toward efficient agnostic learning," *Machine Learning*, vol. 17, no. 2-3, pp. 115–141, 1994.

[59] V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami, "New results for learning noisy parities and halfspaces," in *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE, 2006, pp. 563–574.

[60] V. Guruswami and P. Raghavendra, "Hardness of learning halfspaces with noise," *SIAM Journal on Computing*, vol. 39, no. 2, pp. 742–765, 2009.

[61] A. Daniely, "Complexity theoretic limitations on learning halfspaces," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, 2016, pp. 105–117.

[62] A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio, "Agnostically learning halfspaces," *SIAM Journal on Computing*, vol. 37, no. 6, pp. 1777–1805, 2008.

[63] A. Klivans and P. Kothari, "Embedding hard learning problems into gaussian space," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

[64] I. Diakonikolas, D. M. Kane, and N. Zarifis, "Near-optimal sq lower bounds for agnostically learning halfspaces and relus under gaussian marginals," *arXiv preprint arXiv:2006.16200*, 2020.

[65] S. Goel, A. Gollakota, and A. Klivans, "Statistical-query lower bounds via functional gradients," *arXiv preprint arXiv:2006.15812*, 2020.

[66] A. R. Klivans, P. M. Long, and R. A. Servedio, "Learning halfspaces with malicious noise." *Journal of Machine Learning Research*, vol. 10, no. 12, 2009.

[67] M.-F. Balcan and H. Zhang, "Sample and computationally efficient learning algorithms under s-concave distributions," *arXiv preprint arXiv:1703.07758*, 2017.

[68] S. Frei, Y. Cao, and Q. Gu, "Provable generalization of sgd-trained neural networks of any width in the presence of adversarial label noise," *arXiv preprint arXiv:2101.01152*, 2021.

[69] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[70] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[71] I. Mukherjee, C. Rudin, and R. Schapire, "The convergence rate of AdaBoost," in *COLT*, 2011.

[72] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," *arXiv preprint arXiv:1901.08584*, 2019.

[73] L. Chizat, E. Oyallon, and F. Bach, "On lazy training in differentiable programming," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[74] C. Wei, J. D. Lee, Q. Liu, and T. Ma, "Regularization matters: Generalization and optimization of neural nets vs their induced kernel," *arXiv preprint arXiv:1810.05369*, 2018.

[75] A. Nitanda, G. Chinot, and T. Suzuki, "Gradient descent can learn less over-parameterized two-layer neural networks on classification problems," *arXiv preprint arXiv:1905.09870*, 2019.

[76] S. S. Du, X. Zhai, B. Poczos, and A. Singh, "Gradient descent provably optimizes over-parameterized neural networks," *arXiv preprint arXiv:1810.02054*, 2018.

[77] S. Oymak and M. Soltanolkotabi, "Towards moderate overparameterization: global convergence guarantees for training shallow neural networks," *arXiv preprint arXiv:1902.04674*, 2019.

[78] Z. Song and X. Yang, "Quadratic suffices for over-parametrization via matrix chernoff bound," *arXiv preprint arXiv:1906.03593*, 2019.

[79] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in *Advances in Neural Information Processing Systems*, 2018, pp. 8157–8166.

[80] Z. Allen-Zhu, Y. Li, and Y. Liang, "Learning and generalization in overparameterized neural networks, going beyond two layers," *arXiv preprint arXiv:1811.04918*, 2018.

[81] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," *arXiv preprint arXiv:1811.03962*, 2018.

[82] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," *arXiv preprint arXiv:1811.03804*, 2018.

[83] D. Zou, Y. Cao, D. Zhou, and Q. Gu, "Stochastic gradient descent optimizes over-parameterized deep relu networks," *arXiv preprint arXiv:1811.08888*, 2018.

[84] D. Zou and Q. Gu, "An improved analysis of training over-parameterized deep neural networks," *arXiv preprint arXiv:1906.04688*, 2019.

[85] Z. Allen-Zhu, Y. Li, and Z. Song, "On the convergence rate of training recurrent neural networks," *arXiv preprint arXiv:1810.12065*, 2018.

[86] Z. Allen-Zhu and Y. Li, "Can sgd learn recurrent neural networks with provable generalization?" *arXiv preprint arXiv:1902.01028*, 2019.

[87] Z. Allen-Zhu and Y. Li, "What can resnet learn efficiently, going beyond kernels?" *arXiv preprint arXiv:1905.10337*, 2019.

[88] Y. Cao and Q. Gu, "Generalization bounds of stochastic gradient descent for wide and deep neural networks," *arXiv preprint arXiv:1905.13210*, 2019.

[89] M. J. Wainwright, "UC Berkeley Statistics 210B, Lecture Notes: Basic tail and concentration bounds," Jan 2015. [Online]. Available: https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf

[90] N. Srebro, K. Sridharan, and A. Tewari, "Smoothness, low noise and fast rates," in *Advances in neural information processing systems*, 2010, pp. 2199–2207.

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

[92] P. Liang, "Stanford CS229T/STAT231: Statistical Learning Theory," Apr 2016. [Online]. Available: https://web.stanford.edu/class/cs229t/notes.pdf

[93] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *JMLR*, vol. 3, pp. 463–482, Nov 2002.

[94] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire, "Contextual bandit algorithms with supervised learning guarantees," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 19–26.

[95] J. M. Borwein and Q. J. Zhu, *Techniques of Variational Analysis, volume 20 of*. CMS Books in Mathematics, 2005.

[96] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 9461–9471.

[97] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[98] S. Arora, N. Cohen, and E. Hazan, "On the optimization of deep networks: Implicit acceleration by overparameterization," *arXiv preprint arXiv:1802.06509*, 2018.

[99] H. Lu and K. Kawaguchi, "Depth creates no bad local minima," *arXiv preprint arXiv:1702.08580*, 2017.

[100] T. Laurent and J. von Brecht, "Deep linear neural networks with arbitrary loss: All local minima are global," *arXiv preprint arXiv:1712.01473*, 2017.

[101] Y. Zhou and Y. Liang, "Critical points of linear neural networks: Analytical forms and landscape properties," 2018.

[102] S. S. Du, W. Hu, and J. D. Lee, "Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced," *arXiv preprint arXiv:1806.00900*, 2018.

[103] K. Lyu and J. Li, "Gradient descent maximizes the margin of homogeneous neural networks," *arXiv preprint arXiv:1906.05890*, 2019.

[104] L. Chizat and F. Bach, "Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss," *arXiv preprint arXiv:2002.04486*, 2020.

[105] P. M. Long, "Properties of the after kernel," *arXiv preprint arXiv:2105.10585*, 2021.

Appendix A: Technical lemmas

Here are some technical results needed in this thesis.

Lemma A.1. Let $r, \rho > 0$ be given, then

$$\frac{2}{\rho}(1 - e^{-r\rho}) \leq \int_0^{2\pi} \ell_{\log}\left(r\rho\big|\cos(\theta)\big|\right) r \,\mathrm{d}\theta \leq \frac{8\sqrt{2}}{\rho}.$$

Proof. First note that by symmetry,

$$\int_0^{2\pi} \ell_{\log}\left(r\rho\big|\cos(\theta)\big|\right) r \,\mathrm{d}\theta = 4\int_0^{\frac{\pi}{2}} \ell_{\log}\left(r\rho\cos(\theta)\right) r \,\mathrm{d}\theta.$$

On the upper bound, note that $\ell_{\log}\left(r\rho\cos(\theta)\right)$ is increasing as θ goes from 0 to $\frac{\pi}{2}$, and moreover $\sin(\theta) \geq \frac{\sqrt{2}}{2}$ for $\theta \in \left(\frac{\pi}{4}, \frac{\pi}{2}\right)$, therefore

$$4\int_0^{\frac{\pi}{2}} \ell_{\log}\left(r\rho\cos(\theta)\right) r \,\mathrm{d}\theta \leq 8\int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \ell_{\log}\left(r\rho\cos(\theta)\right) r \,\mathrm{d}\theta \leq \frac{8\sqrt{2}}{\rho}\int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \ell_{\log}\left(r\rho\cos(\theta)\right) r\rho\sin(\theta) \,\mathrm{d}\theta.$$

Also because $\ell_{\log}(z) \leq \exp(-z)$,

$$\begin{aligned}
\int_0^{2\pi} \ell_{\log}\left(r\rho\big|\cos(\theta)\big|\right) r \,\mathrm{d}\theta &\leq \frac{8\sqrt{2}}{\rho}\int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \exp\left(-r\rho\cos(\theta)\right) r\rho\sin(\theta) \,\mathrm{d}\theta \\
&= \frac{8\sqrt{2}}{\rho}\left(1 - \exp\left(-\frac{\sqrt{2}r\rho}{2}\right)\right) \\
&\leq \frac{8\sqrt{2}}{\rho}.
\end{aligned}$$

On the lower bound, note that $\ell_{\log}(z) \geq \frac{1}{2}\exp(-z)$ for $z \geq 0$, therefore

$$\begin{aligned}
\int_0^{2\pi} \ell_{\log}\left(r\rho\big|\cos(\theta)\big|\right) r \,\mathrm{d}\theta &= 4\int_0^{\frac{\pi}{2}} \ell_{\log}\left(r\rho\cos(\theta)\right) r \,\mathrm{d}\theta \\
&\geq 2\int_0^{\frac{\pi}{2}} \exp\left(-r\rho\cos(\theta)\right) r \,\mathrm{d}\theta \\
&\geq \frac{2}{\rho}\int_0^{\frac{\pi}{2}} \exp\left(-r\rho\cos(\theta)\right) r\rho\sin(\theta) \,\mathrm{d}\theta \\
&= \frac{2}{\rho}\left(1 - e^{-r\rho}\right).
\end{aligned}$$

QED.

Lemma A.2. Given $w, w' \in \mathbb{R}^d$, suppose $\Pr_{(x,y)\sim P}\Big(y \neq \operatorname{sign}\big(\langle w, x\rangle\big)\Big) = \mathrm{OPT}$. If $\|x\|_2 \le B$ almost surely, then

$$\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq \operatorname{sign}\left(\langle w,x\rangle\right)}\big|\langle w', x\rangle\big|\right] \le B\|w'\|_2 \cdot \mathrm{OPT}.$$

If P_x is (α_1, α_2)-sub-exponential, and $\mathrm{OPT} \le \frac{1}{e}$, then

$$\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq \operatorname{sign}\left(\langle w,x\rangle\right)}\big|\langle w', x\rangle\big|\right] \le (1+2\alpha_1)\alpha_2\|w'\|_2 \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right).$$

Proof. If $\|x\|_2 \le B$ almost surely, then

$$\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq \operatorname{sign}\left(\langle w,x\rangle\right)}\big|\langle w', x\rangle\big|\right] \le B\|w'\|_2\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq \operatorname{sign}\left(\langle w,x\rangle\right)}\right] = B\|w'\|_2 \cdot \mathrm{OPT}.$$

Below we assume P_x is (α_1, α_2)-sub-exponential.

Let $\nu_x := \langle w', x\rangle$; we first give some tail bounds for ν_x. Since P_x is (α_1, α_2)-sub-exponential, for any $t > 0$, we have

$$\Pr\left(\left|\left\langle \frac{w'}{\|w'\|_2}, x\right\rangle\right| \ge t\right) \le \alpha_1 \exp\left(-\frac{t}{\alpha_2}\right), \quad \text{or} \quad \Pr\big(|\nu_x| \ge t\big) \le \alpha_1 \exp\left(-\frac{t}{\alpha_2\|w'\|_2}\right).$$

Let $\mu(t) := \Pr\big(|\nu_x| \ge t\big)$. Given any threshold $\tau > 0$, integration by parts gives

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{|\nu_x|\ge\tau}|\nu_x|\right] &= \int_\tau^\infty t \cdot \big(-\,\mathrm{d}\mu(t)\big) \\ &= \tau\mu(\tau) + \int_\tau^\infty \mu(t)\,\mathrm{d}t \le \alpha_1\left(\alpha_2\|w'\|_2 + \tau\right)\exp\left(-\frac{\tau}{\alpha_2\|w'\|_2}\right). \end{aligned} \tag{A.1}$$

Now let $\tau := \alpha_2\|w'\|_2 \ln\left(\frac{1}{\mathrm{OPT}}\right)$. Note that

$$\begin{aligned}\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq \operatorname{sign}\left(\langle w,x\rangle\right)}\big|\langle w', x\rangle\big|\right] = \mathbb{E}_{(x,y)\sim P}&\left[\mathbb{1}_{|\nu_x|\le\tau}\mathbb{1}_{y\neq \operatorname{sign}\left(\langle w,x\rangle\right)}|\nu_x|\right] \\ &+ \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{|\nu_x|\ge\tau}\mathbb{1}_{y\neq \operatorname{sign}\left(\langle w,x\rangle\right)}|\nu_x|\right].\end{aligned}$$

We bound the two parts separately. When $|\nu_x| \le \tau$, we have

$$\mathbb{E}\left[\mathbb{1}_{|\nu_x|\le\tau}\mathbb{1}_{y\neq \operatorname{sign}\left(\langle w,x\rangle\right)}|\nu_x|\right] \le \tau\mathbb{E}\left[\mathbb{1}_{y\neq \operatorname{sign}\left(\langle w,x\rangle\right)}\right] = \tau \cdot \mathrm{OPT} = \alpha_2\|w'\|_2 \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right).$$

On the other hand, when $|\nu_x| \geq \tau$, eq. (A.1) gives

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{|\nu_x|\geq\tau}\mathbb{1}_{y\neq\operatorname{sign}\left(\langle w,x\rangle\right)}|\nu_x|\right] &\leq \mathbb{E}\left[\mathbb{1}_{|\nu_x|\geq\tau}|\nu_x|\right] \\
&\leq \alpha_1\alpha_2\|w'\|_2\left(1+\ln\left(\frac{1}{\mathrm{OPT}}\right)\right)\mathrm{OPT} \\
&\leq 2\alpha_1\alpha_2\|w'\|_2\cdot\mathrm{OPT}\cdot\ln\left(\frac{1}{\mathrm{OPT}}\right),
\end{aligned}$$

where we also use $\mathrm{OPT} \leq \frac{1}{e}$. To sum up,

$$\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq\operatorname{sign}\left(\langle w,x\rangle\right)}\left|\langle w',x\rangle\right|\right] \leq (1+2\alpha_1)\alpha_2\|w'\|_2\cdot\mathrm{OPT}\cdot\ln\left(\frac{1}{\mathrm{OPT}}\right).$$

QED.