Agnostic Learning of Halfspaces with Gradient Descent via Soft Margins

Spencer Frei [1] **Yuan Cao** [2] **Quanquan Gu** [2]

Abstract

We analyze the properties of gradient descent on convex surrogates for the zero-one loss for the agnostic learning of halfspaces. We show that when a quantity we refer to as the *soft margin* is well-behaved—a condition satisfied by log-concave isotropic distributions among others—minimizers of convex surrogates for the zero-one loss are approximate minimizers for the zero-one loss itself. As standard convex optimization arguments lead to efficient guarantees for minimizing convex surrogates of the zero-one loss, our methods allow for the first positive guarantees for the classification error of halfspaces learned by gradient descent using the binary cross-entropy or hinge loss in the presence of agnostic label noise.

1. Introduction

We analyze the performance of gradient descent on a convex surrogate for the zero-one loss in the context of the agnostic learning of halfspaces. By a *halfspace* we mean a function $x \mapsto \mathrm{sgn}(w^\top x) \in \{\pm 1\}$ for some $w \in \mathbb{R}^d$. Let $\mathcal{D}$ be a joint distribution over (x, y), where the inputs $x \in \mathbb{R}^d$ and the labels $y \in \{\pm 1\}$, and denote by $\mathcal{D}_x$ the marginal of $\mathcal{D}$ over x. We are interested in the performance of halfspaces found by gradient descent in comparison to the best-performing halfspace over $\mathcal{D}$, so let us define, for $w \in \mathbb{R}^d$,

$$\mathrm{err}_{\mathcal{D}}^{0-1}(w) := \mathbb{P}_{(x,y)\sim\mathcal{D}}(\mathrm{sgn}(w^\top x) \neq y),$$
$$\mathsf{OPT} := \min_{\|w\|=1} \mathrm{err}_{\mathcal{D}}^{0-1}(w).$$

We consider the *agnostic* setting, i.e. we make no assumptions on the relationship between x and y and so in general $\mathsf{OPT} > 0$. Due to the non-convexity and discontinuity of the zero-one loss, the standard approach for minimizing the classification error is to consider a convex surrogate loss $\ell : \mathbb{R} \to \mathbb{R}$ for which $\mathbb{1}(z < 0) \leq O(\ell(z))$ and to instead minimize the surrogate risk

$$F_\ell(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(yw^\top x)\big]. \tag{1}$$

Without access to the population risk itself, one can take samples $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ and optimize (1) by gradient descent on the empirical risk $\widehat{F_\ell}(w)$, defined by taking the expectation in (1) over the empirical distribution of the samples. By using standard tools from convex optimization and Rademacher complexity, such an approach is guaranteed to efficiently minimize the population surrogate risk up to optimization and statistical error. The question is then, given that we have found a halfspace $x \mapsto w^\top x$ that minimizes the *surrogate* risk, how does this halfspace compare to the *best* halfspace as measured by the zero-one loss? And how does the choice of the surrogate loss affect this behavior? To the best of our knowledge, no previous work has been able to demonstrate that gradient descent on convex surrogates can yield approximate minimizers for the classification error over halfspaces, even for the case of the standard logistic (binary cross-entropy) loss $\ell(z) = \log(1 + \exp(-z))$ or the hinge loss $\ell(z) = \max(1 - z, 0)$.

We show below that the answer to these questions depend upon what we refer to as the *soft margin function* of the distribution at a given minimizer for the zero-one loss. (We note that in general, there may be multiple minimizers for the zero-one loss, and so we can only refer to *a* given minimizer.) For $\bar{v} \in \mathbb{R}^d$ satisfying $\|\bar{v}\| = 1$, we say that the halfspace $\bar{v}$ satisfies the $\phi_{\bar{v}}$-soft-margin property if for some function $\phi_{\bar{v}} : [0, 1] \to \mathbb{R}$, for all $\gamma \in [0, 1]$,

$$\mathbb{P}_{\mathcal{D}_x}(|\bar{v}^\top x| \leq \gamma) \leq \phi_{\bar{v}}(\gamma).$$

The key insight of our analysis is that the soft margin can be used to bound the error under *convex surrogates* for the zero-one loss by the error achieved under the zero-one loss itself. In particular, for bounded distributions $\mathcal{D}_x$, we show in Lemma 5.1 below that

$$F_\ell(v) \leq \inf_{\gamma>0}\Big\{(1 + LB_X\gamma^{-1}\ell^{-1}(\varepsilon))\mathsf{OPT} + \phi(\gamma) + \varepsilon\Big\}.$$

where $\phi_{\bar{v}}$ is a soft margin function corresponding to a unit norm minimizer $\bar{v}$ of the population zero-one loss, and v is a scalar multiple of $\bar{v}$. Thus, provided $\phi_{\bar{v}}(\gamma)$ is well-behaved in the sense that $\phi_{\bar{v}}(\gamma)$ is small when γ is small,

[1]Department of Statistics, UCLA [2]Department of Computer Science, UCLA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

minimizers of convex surrogates for the zero-one loss will be approximate minimizers for the zero-one loss itself. This implies that any black-box optimization algorithm which can efficiently minimize convex functions, like gradient descent on the logistic loss, will produce halfspaces which are approximate minimizers for the zero-one loss itself. In particular, we are able to show the following guarantees for the output of gradient descent on convex surrogates for the zero-one loss.

1. **Hard margin distributions**. If $\|x\| \leq B_X$ almost surely and there is $\bar{\gamma} > 0$ such that $\bar{v}^\top x \geq \bar{\gamma}$ a.s., then $\mathrm{err}_{\mathcal{D}}^{0-1}(w_t) \leq \tilde{O}(\bar{\gamma}^{-1}\mathsf{OPT}) + \varepsilon$.
2. **Sub-exponential distributions satisfying anti-concentration.** If random vectors from $\mathcal{D}_x$ are sub-exponential and satisfy an anti-concentration inequality for projections onto one dimensional subspaces, then $\mathrm{err}_{\mathcal{D}}^{0-1}(w_t) \leq \tilde{O}(\mathsf{OPT}^{1/2}) + \varepsilon$. This covers any log-concave isotropic distribution.

For each of our guarantees, the runtime and sample complexity are $\mathrm{poly}(d, \varepsilon^{-1})$. The exact rates are given in Corollaries 5.3, 5.6 and 5.11. In Table 1 we compare our results with known lower bounds in the literature. To the best of our knowledge, our results are the first to show that gradient descent on convex surrogates for the zero-one loss can learn halfspaces in the presence of agnostic label noise, despite the ubiquity of this approach for classification problems.

The remainder of the paper is organized as follows. In Section 2, we review the literature on learning halfspaces in the presence of noise. In Section 3, we discuss the notion of soft margins which will be essential to our proofs, and provide examples of soft margin behavior for different distributions. In Section 4 we show that gradient descent efficiently finds minimizers of convex surrogate risks and discuss how the tail behavior of the loss function can affect the time and sample complexities of gradient descent. In Section 5 we provide our main results, which relies upon using soft margins to convert minimizers for the convex surrogate risk to approximate minimizers for the classification error. We conclude in Section 6.

2. Related Work

The problem of learning halfspaces is a classical problem in machine learning with a history almost as long as the history of machine learning itself, starting from the perceptron (Rosenblatt, 1958) and support vector machines (Boser et al., 1992) to today. Much of the early works on this problem focused on the realizable setting, i.e. where $\mathsf{OPT} = 0$. In this setting, the Perceptron algorithm or methods from linear programming can be used to efficiently find the optimal halfspace. In the setting of agnostic PAC learning (Kearns et al., 1994) where $\mathsf{OPT} > 0$ in general, the question of which distributions can be learned up to classification error $\mathsf{OPT} + \varepsilon$, and whether it is possible to do so in $\mathrm{poly}(d, 1/\varepsilon)$ time (where d is the input dimension), is significantly more difficult and is still an active area of research. It is known that without distributional assumptions, learning up to even $O(\mathsf{OPT}) + \varepsilon$ is NP-hard, both for proper learning (Guruswami & Raghavendra, 2009) and improper learning (Daniely, 2016). Due to this difficulty, it is common to make a number of assumptions on either $\mathcal{D}_x$ or to impose some type of structure to the learning problem.

A common structure imposed is that of structured noise: one can assume that there exists some underlying halfspace $y = \mathrm{sgn}(v^\top x)$ that is corrupted with probability $p(x) \in [0, 1]$, possibly dependent on the features x. The simplest setting is that of random classification noise, where $p(x) \equiv \eta$, so that each label is flipped with the same probability (Angluin & Laird, 1988); polynomial time algorithms for learning under this noise condition were shown by Blum et al. (1998). The Massart noise model introduced by Massart et al. (2006) relaxes this assumption to $p(x) \leq p$ for some absolute constant $p < 1/2$. The Tsybakov noise model (Tsybakov et al., 2004) is a generalization of the Massart noise model that instead requires a tail bound on $\mathbb{P}(p(x) \geq 1/2 - t)$ for $t > 0$. Awasthi et al. (2015) showed that optimally learning halfspaces under Massart noise is possible for the uniform distribution on the unit sphere, and Awasthi et al. (2016) showed this for log-concave isotropic distributions. The recent landmark result of Diakonikolas et al. (2019) provided the first distribution-independent result for optimally learning halfspaces under Massart noise, answering a long-standing (Sloan, 1988) open problem in computational learning.

By contrast, in the agnostic PAC learning setting, one makes no assumptions on $p(x)$, so one can equivalently view agnostic PAC learning as an adversarial noise model in which an adversary can corrupt the label of a sample x with any probability $p(x) \in [0, 1]$. Recent work suggests that even when $\mathcal{D}_x$ is the Gaussian, agnostically learning up to exactly $\mathsf{OPT} + \varepsilon$ likely requires $\exp(1/\varepsilon)$ time (Goel et al., 2020; Diakonikolas et al., 2020b). In terms of positive results in the agnostic setting, Kalai et al. (2008) showed that a variant of the Average algorithm (Servedio, 1999) can achieve risk $O(\mathsf{OPT}\sqrt{\log(1/\mathsf{OPT})})$ risk in $\mathrm{poly}(d, 1/\varepsilon)$ time when $\mathcal{D}_x$ is uniform over the unit sphere. Awasthi et al. (2017) demonstrated that a localization-based algorithm can achieve $O(\mathsf{OPT}) + \varepsilon$ under log-concave isotropic marginals. Diakonikolas et al. (2020d) showed that for a broad class of distributions, the output of projected SGD on a nonconvex surrogate for the zero-one loss produces a halfspace with risk $O(\mathsf{OPT}) + \varepsilon$ in $\mathrm{poly}(d, 1/\varepsilon)$ time. For more background on learning halfspaces in the presence of noise, we refer the reader to Balcan & Haghtalab (2021).

Table 1. Comparison of our results with other upper and lower bounds in the literature.

Algorithm	$\mathcal{D}_x$	Population Risk	Known Lower Bound
Non-convex G.D. (Diakonikolas et al., 2020d)	Concentration, anti-concentration	$O(\mathsf{OPT})$	N/A
Convex G.D. (this paper)	Sub-exponential, anti-concentration	$\tilde{O}(\mathsf{OPT}^{1/2})$	$\Omega(\mathsf{OPT} \cdot \mathrm{polylog}(1/\mathsf{OPT}))$ (Diakonikolas et al., 2020d)
Convex G.D. (this paper)	s-heavy tail ($s > 2$), anti-concentration	$\tilde{O}(\mathsf{OPT}^{s/(2+2s)})$	$\Omega(\mathsf{OPT}^{1-1/s})$ (Diakonikolas et al., 2020d)
Convex G.D. (this paper)	Hard margin	$\tilde{O}(\bar{\gamma}^{-1}\mathsf{OPT})$	$\Omega(\bar{\gamma}^{-1}\mathsf{OPT})$ (Diakonikolas et al., 2019)

We note that Diakonikolas et al. (2020d) also showed that the minimizer of the surrogate risk of any *convex* surrogate for the zero-one loss is a halfspace with classification error $\omega(\mathsf{OPT})$. Ben-David et al. (2012) and Awasthi et al. (2017) showed similar lower bounds that together imply that empirical risk minimization procedures for convex surrogates yield halfspaces with classification error $\Omega(\mathsf{OPT})$. Given such lower bounds, we wish to emphasize that in this paper we are *not* making a claim about the optimality of gradient descent (on convex surrogates) for learning halfspaces. Rather, our main interest is the characterization of what are the strongest learning guarantees possible with what is perhaps the simplest learning algorithm possible. Given the success of gradient descent for the learning of deep neural networks, and the numerous questions that this success has brought to the theory of statistics and machine learning, we think it is important to develop a thorough understanding of what are the possibilities of vanilla gradient descent, especially in the simplest setting possible.

Recent work has shown that gradient descent finds approximate minimizers for the population risk of single neurons $x \mapsto \sigma(w^\top x)$ under the squared loss (Diakonikolas et al., 2020a; Frei et al., 2020), despite the computational intractability of finding the optimal single neuron (Goel et al., 2019). The main contribution of this paper is that despite the computational difficulties in *exact* agnostic learning, the standard gradient descent algorithm satisfies an *approximate* agnostic PAC learning guarantee, in line with the results found by Frei et al. (2020) for the single neuron.

2.1. Notation

We say that a differentiable loss function ℓ is L-Lipschitz if $|\ell'(z)| \leq L$ for all z in its domain, and we say the loss is H-smooth if its derivative ℓ' is H-Lipschitz. We use the word "decreasing" interchangeably with "non-increasing". We use the standard $O(\cdot), \Omega(\cdot)$ order notations to hide universal constants and $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ to additionally suppress logarithmic factors. Throughout this paper, $\|x\|$ refers to the standard Euclidean norm on $\mathbb{R}^d$ induced by the inner product $x^\top x$. We will emphasize that a vector v is of unit norm by writing $\bar{v}$. We assume $\mathcal{D}$ is a probability distribution over $\mathbb{R}^d \times \{\pm 1\}$ with marginal distribution $\mathcal{D}_x$ over $\mathbb{R}^d$. For general decreasing function ℓ, for which an inverse function may or may not exist, we overload the notation ℓ^{-1} by denoting $\ell^{-1}(t) := \inf\{z : \ell(z) \leq t\}$.

3. Soft Margins

In this section we will formally introduce the soft margin function and describe some common distributions for which it takes a simple form.

Definition 3.1. *Let $\bar{v} \in \mathbb{R}^d$ satisfy $\|\bar{v}\| = 1$. We say $\bar{v}$ satisfies the* soft margin condition *with respect to a function $\phi_{\bar{v}} : \mathbb{R} \to \mathbb{R}$ if for all $\gamma \in [0, 1]$, it holds that*

$$\mathbb{E}_{x \sim \mathcal{D}_x}\left[\mathbb{1}\left(|\bar{v}^\top x| \leq \gamma\right)\right] \leq \phi_{\bar{v}}(\gamma).$$

We note that our definition of soft margin is essentially an unnormalized version of the soft margin function considered by Foster et al. (2018) in the context of learning GLMs, since they defined $\phi_{\bar{v}}(\gamma)$ as the probability that $|\bar{v}^\top x / \|x\|| \leq \gamma$. This concept was also considered by Balcan & Zhang (2017) for s-concave isotropic distributions under the name 'probability of a band'.

Below we will consider some examples of soft margin function behavior. We shall see later that our final generalization bounds will depend on the behavior of $\phi_{\bar{v}}(\gamma)$ for γ sufficiently small, and thus in the below examples we only care about the behavior of $\phi_{\bar{v}}(\cdot)$ in small neighborhoods of the origin. In our first example, we show that (hard) margin distributions have simple soft margin functions.

Example 3.2 (Hard margin distributions)**.** *If $\mathcal{D}_x$ is a hard margin distribution in the sense that $\bar{v}^\top x \geq \gamma^* > 0$ for some $\gamma^* > 0$ almost surely, then $\phi_{\bar{v}}(\gamma) = 0$ for $\gamma < \gamma^*$.*

Proof. This follows immediately: $\mathbb{P}(|\bar{v}^\top x| \leq \gamma) = 0$ when $\gamma < \gamma^*$. □

Note that the soft margin function in Example 3.2 is specific to the vector $\bar{v}$, and does not necessarily hold for arbitrary unit vectors in $\mathbb{R}^d$. By contrast, for many distributions it is possible to derive bounds on soft margin functions that hold for *any* vector $\bar{v}$, which we shall see below is a key step for deriving approximate agnostic learning guarantees for the output of gradient descent.

The next example shows that provided the projections of $\mathcal{D}_x$ onto one dimensional subspaces satisfy an anti-concentration property, then all soft margins function for that distribution take a simple form. To do so we first introduce the following definition.

Definition 3.3 (Anti-concentration)**.** *For $\bar{v} \in \mathbb{R}^d$, denote by $p_{\bar{v}}(\cdot)$ the marginal distribution of $x \sim \mathcal{D}_x$ on the subspace spanned by $\bar{v}$. We say $\mathcal{D}_x$ satisfies U-*anti-concentration *if there is some $U > 0$ such that for all unit norm $\bar{v}$, $p_{\bar{v}}(z) \leq U$ for all $z \in \mathbb{R}$.*

A similar assumption was used in Diakonikolas et al. (2020c;d;e) for learning halfspaces; in their setup, the anti-concentration assumption was for the projections of $\mathcal{D}_x$ onto two dimensional subspaces rather than the one dimensional version we consider here.

Example 3.4 (Distributions satisfying anti-concentration)**.** *If $\mathcal{D}_x$ satisfies U-anti-concentration, then for any unit norm $\bar{v}$, $\phi_{\bar{v}}(\gamma) \leq 2U\gamma$.*

Proof. We can write $\mathbb{P}(|\bar{v}^\top x| \leq \gamma) = \int_{-\gamma}^{\gamma} p_{\bar{v}}(z)\mathrm{d}z \leq 2\gamma U$. □

We will show below that log-concave isotropic distributions satisfy U-anti-concentration for $U = 1$. We first remind the reader of the definition of log-concave isotropic distributions.

Definition 3.5. *We say that a distribution $\mathcal{D}_x$ over $x \in \mathbb{R}^d$ is* log-concave *if it has a density function $p(\cdot)$ such that $\log p(\cdot)$ is concave. We call $\mathcal{D}_x$* isotropic *if its mean is the zero vector and its covariance matrix is the identity matrix.*

Typical examples of log-concave isotropic distributions include the standard Gaussian and the uniform distribution over a convex set.

Example 3.6 (Log-concave isotropic distributions)**.** *If $\mathcal{D}_x$ is log-concave isotropic then it satisfies 1-anti-concentration, and thus for any $\bar{v}$ with $\|\bar{v}\| = 1$, $\phi_{\bar{v}}(\gamma) \leq 2\gamma$.*

Proof. This was demonstrated in Balcan & Zhang (2017, Proof of Theorem 11).[1] □

[1]The cited theorem implies a similar bound of the form $O(\gamma)$ holds for the more general set of s-concave isotropic distributions. We focus here on log-concave isotropic distributions for simplicity.

4. Gradient Descent Finds Minimizers of the Surrogate Risk

We begin by demonstrating that gradient descent finds weights that achieve the best population-level surrogate risk. The following theorem is a standard result from stochastic optimization. For completeness, we present its proof in Appendix E.

Theorem 4.1. *Suppose $\|x\| \leq B_X$ a.s. Let ℓ be convex, L-Lipschitz, and H-smooth, with $\ell(0) \leq 1$. Let $v \in \mathbb{R}^d$ be arbitrary with $\|v\| \leq V$ for some $V > 1$, and suppose that the initialization w_0 satisfies $\|w_0\| \leq V$. For any $\varepsilon, \delta > 0$ and for any provided $\eta \leq (2/5)H^{-1}B_X^{-2}$, if gradient descent is run for $T = (4/3)\eta^{-1}\varepsilon^{-1}\|w_0 - v\|^2$, then with probability at least $1 - \delta$,*

$$F_\ell(w_{T-1}) \leq F_\ell(v) + \frac{4B_X V L}{\sqrt{n}} + 8B_X V\sqrt{\frac{2\log(2/\delta)}{n}} + \varepsilon.$$

This shows that gradient descent learns halfspaces that have a population surrogate risk competitive with that of the best predictor with bounded norm for any norm threshold V. For distributions that are linearly separable by some margin $\gamma > 0$, the above theorem allows us to derive upper bounds on the sample complexity that suggest that exponentially tailed losses are preferable to polynomially tailed losses from both time and sample complexity perspectives, touching on a recent problem posed by Ji et al. (2020).

Corollary 4.2 (Sample complexity for linearly separable data)**.** *Assume $\|x\| \leq B_X$ a.s. Suppose that for some $\bar{v} \in \mathbb{R}^d$, $\|\bar{v}\| = 1$, there is $\gamma > 0$ such that $y\bar{v}^\top x \geq \gamma$ a.s. If ℓ is convex, decreasing, L-Lipschitz, and H-smooth, and if we fix a step size of $\eta \leq (2/5)H^{-1}B_X^{-2}$, then*

- *Assume ℓ has polynomial tails, so that for some $C_0, p > 0$ and $\ell(z) \leq C_0 z^{-p}$ holds for all $z \geq 1$. Provided $n = \Omega(\gamma^{-2}\varepsilon^{-2-2/p})$, then running gradient descent for $T = \Omega(\gamma^{-2}\varepsilon^{-1-2/p})$ iterations guarantees that $\mathrm{err}_{\mathcal{D}}^{0-1}(w_T) \leq \varepsilon$.*
- *Assume ℓ has exponential tails, so that for some $C_0, C_1, p > 0$, $\ell(z) \leq C_0 \exp(-C_1 z^p)$ holds for all $z \geq 1$. Then $n = \tilde{\Omega}(\gamma^{-2}\varepsilon^{-2})$ and $T = \tilde{\Omega}(\gamma^{-2}\varepsilon^{-1})$ guarantees that $\mathrm{err}_{\mathcal{D}}^{0-1}(w_T) \leq \varepsilon$.*

The proof for the above Corollary can be found in Appendix D. At a high level, the above result shows that if the tails of the loss function are heavier, one may need to run gradient

descent for longer to drive the population surrogate risk, and hence the zero-one risk, to zero.[2] In the subsequent sections, we shall see that this phenomenon persists beyond the linearly separable case to the more general agnostic learning setting.

Remark 4.3. *The sample complexity in Theorem 4.1 can be improved from $O(\varepsilon^{-2})$ to $O(\varepsilon^{-1})$ if we use online stochastic gradient descent rather than vanilla gradient descent. The proof of this is somewhat more involved as it requires a technical workaround to the unboundedness of the loss function, and may be of independent interest. We present the full analysis of this in Appendix A.*

5. Gradient Descent Finds Approximate Minimizers for the Zero-One Loss

We now show how we can use the soft margin function to develop bounds for the zero-one loss of the output of gradient descent.

5.1. Bounded Distributions

We first focus on the case when the marginal distribution $\mathcal{D}_x$ is bounded almost surely.

By Theorem 4.1, since by Markov's inequality we have that $\mathrm{err}_{\mathcal{D}}^{0-1}(w) \le \ell(0)^{-1} F_\ell(w)$, if we want to show that the zero-one population risk for the output of gradient descent is competitive with that of the optimal zero-one loss achieved by some halfspace $v \in \mathbb{R}^d$, it suffices to bound $F_\ell(v)$ by some function of OPT. To do so we decompose the expectation for $F_\ell(v)$ into a sum of three terms which incorporate OPT, the soft margin function, and a term that drives the surrogate risk to zero by driving up the margin on those samples that are correctly classified.

Lemma 5.1. *Let $\bar{v}$ be a unit norm population risk minimizer for the zero-one loss, and suppose $\bar{v}$ satisfies the soft margin condition with respect to some $\phi : [0,1] \to \mathbb{R}$. Assume that $\|x\| \le B_X$ a.s. Let $v = V\bar{v}$ for $V > 0$ be a scaled version of $\bar{v}$. If ℓ is decreasing, L-Lipschitz and $\ell(0) \le 1$, then*

$$F_\ell(v) \le \inf_{\gamma > 0} \left\{ (1 + LVB_X)\mathsf{OPT} + \phi(\gamma) + \ell(V\gamma) \right\}.$$

In particular, for $v = \gamma^{-1}\ell^{-1}(\varepsilon)\bar{v}$ for some $\varepsilon > 0$, we have

$$F_\ell(v) \le \inf_{\gamma > 0} \left\{ (1 + LB_X\gamma^{-1}\ell^{-1}(\varepsilon))\mathsf{OPT} + \phi(\gamma) + \varepsilon \right\}.$$

Proof. We begin by writing the expectation as a sum of three terms,

$$\begin{aligned}\mathbb{E}[\ell(yv^\top x)] &= \mathbb{E}\left[\ell(yv^\top x)\mathbb{1}\left(y\bar{v}^\top x \le 0\right)\right] \\ &\quad + \mathbb{E}\left[\ell(yv^\top x)\mathbb{1}\left(0 < y\bar{v}^\top x \le \gamma\right)\right] \\ &\quad + \mathbb{E}\left[\ell(yv^\top x)\mathbb{1}\left(y\bar{v}^\top x > \gamma\right)\right]. \qquad (2)\end{aligned}$$

For the first term, we use that ℓ is L-Lipschitz and decreasing as well as Cauchy–Schwarz to get

$$\begin{aligned}\mathbb{E}[\ell(yv^\top x)\mathbb{1}(y\bar{v}^\top x \le 0)] &\le \mathbb{E}[(1 + L|v^\top x|)\mathbb{1}(y\bar{v}^\top x \le 0)] \\ &\le (1 + LVB_X)\mathbb{E}[\mathbb{1}(y\bar{v}^\top x \le 0)] \\ &= (1 + LVB_X)\mathsf{OPT}.\end{aligned}$$

In the last inequality we use that $\|x\| \le B_X$ a.s. For the second term,

$$\begin{aligned}&\mathbb{E}\left[\ell(yv^\top x)\mathbb{1}\left(0 < y\bar{v}^\top x \le \gamma\right)\right] \\ &\le \ell(0)\mathbb{E}\left[\mathbb{1}\left(0 < y\bar{v}^\top x \le \gamma\right)\right] \le \phi(\gamma), \qquad (3)\end{aligned}$$

where we have used that ℓ is decreasing in the first inequality and Definition 3.1 in the second. Finally, for the last term, we can use that ℓ is decreasing to get

$$\begin{aligned}&\mathbb{E}\left[\ell(yv^\top x)\mathbb{1}\left(y\bar{v}^\top x > \gamma\right)\right] \\ &= \mathbb{E}\left[\ell(yV\bar{v}^\top x)\mathbb{1}\left(yV\bar{v}^\top x > V\gamma\right)\right] \le \ell(V\gamma). \qquad (4)\end{aligned}$$

The final claim comes from taking $V = \gamma^{-1}\ell^{-1}(\varepsilon)$. □

Note that for the hinge loss, $\ell^{-1}(\varepsilon) = 0$ for $\varepsilon \le 1$, while for losses with exponential tails like the binary cross-entropy loss, $\ell^{-1}(\varepsilon) = O(\log(1/\varepsilon))$. This means that the contribution of the $\ell^{-1}(\varepsilon)$ term in Lemma 5.1 is negligible for the losses used in standard black-box optimization algorithms. Thus, Lemma 5.1 shows that we can bound the population risk under convex surrogates of the zero-one loss by a quantity involving OPT, the classification error achieved by minimizers of the zero-one loss, the soft margin $\phi(\gamma)$, and some negligible additional terms. Since gradient descent is able to efficiently minimize the population risk over any norm-boudned domain, we can easily translate this into a guarantee for the weights found by gradient descent, as given in our next theorem.

Theorem 5.2. *Suppose $\|x\| \le B_X$ a.s. Let ℓ be convex, decreasing, L-Lipschitz, and H-smooth, with $0 < \ell(0) \le 1$. Assume that a unit norm population risk minimizer of the zero-one loss, $\bar{v}$, satisfies the ϕ-soft-margin condition for some increasing $\phi : \mathbb{R} \to \mathbb{R}$. Fix a step size $\eta \le (2/5)H^{-1}B_X^{-2}$. Let $\varepsilon_1, \gamma > 0$ and $\varepsilon_2 \ge 0$ be arbitrary. Denote by w_T the output of gradient descent run for $T = (4/3)\eta^{-1}\varepsilon_1^{-1}\gamma^{-2}[\ell^{-1}(\varepsilon_2)]^{-2}$ iterations after initialization at the origin. Then, with probability at least $1 - \delta$,*

$$\begin{aligned}\mathrm{err}_{\mathcal{D}}^{0-1}(w_T) \le {}& {}^1\!/_{\ell(0)}\big[(1 + LB_X\gamma^{-1}\ell^{-1}(\varepsilon_2))\mathsf{OPT} \\ &+ \phi(\gamma) + O(\gamma^{-1}\ell^{-1}(\varepsilon_2)n^{-1/2}) + \varepsilon_1 + \varepsilon_2\big],\end{aligned}$$

[2] We note that in Corollary 4.2, there is a gap for the sample complexity and runtime when using polynomially tailed vs. exponentially tailed losses. However, such a gap may be an artifact of our analysis. Deriving matching lower bounds for the sample complexity or runtime of gradient descent on polynomially tailed losses remains an open problem.

where $O(\cdot)$ hides absolute constants that depend on L, H, and $\log(1/\delta)$.

Proof. We take $v = V\bar{v}$ for a given unit-norm zero-one population risk minimizer $\bar{v}$ in Theorem 4.1 to get that for some universal constant $C > 0$ depending only on L and $\log(1/\delta)$, with probability at least $1 - \delta$,

$$F_\ell(w_T) \le F_\ell(v) + \varepsilon_1/2 + CVB_X n^{-1/2}. \tag{5}$$

By Lemma 5.1, for any $\gamma > 0$ it holds that

$$F_\ell(v) \le (1 + LVB_X)\mathsf{OPT} + \phi(\gamma) + \ell(V\gamma).$$

Again we take $V = \gamma^{-1}\ell^{-1}(\varepsilon_2)$ to get

$$F_\ell(w_T) \le (1 + L\gamma^{-1})\mathsf{OPT} + \phi(\gamma) + O(\gamma^{-1}\ell^{-1}(\varepsilon_2)n^{-1/2}) + \varepsilon_1 + \varepsilon_2. \tag{6}$$

Finally, by Markov's inequality,

$$\mathbb{P}(yw_T^\top x < 0) \le \frac{\mathbb{E}[\ell(yw_{T-1}^\top x)]}{\ell(0)} = \frac{F_\ell(w_T)}{\ell(0)}. \tag{7}$$

Putting (6) together with (7) completes the proof. □

A few comments on the proof of the above theorem are in order. Note that the only place we use smoothness of the loss function is in showing that gradient descent minimizes the population risk in (5), and it is not difficult to remove the H-smoothness assumption to accommodate e.g. the hinge loss; indeed, in Theorem 5.10 below, we provide an analogous result for unbounded distributions using SGD that allows for non-smooth convex surrogates at the cost of requiring a small step size. On the other hand, that ℓ is L-Lipschitz is key to the proof of Lemma 5.1. Non-Lipschitz losses such as the exponential loss or squared hinge loss would incur additional factors of γ^{-1} in front of OPT in the final bound for Theorem 5.2.[3] We shall see below in the proof of Proposition 5.5 that this would yield worse guarantees for $\mathrm{err}_{\mathcal{D}}^{0-1}(w_T)$.

Additionally, in concordance with the result from Corollary 4.2, we see that if the tail of ℓ is fatter, then $\ell^{-1}(\varepsilon_2)$ will be larger and so our guarantees would be worse. In particular, for losses with exponential tails, $\ell^{-1}(\varepsilon_2) = O(\log(1/\varepsilon_2))$, and so by using such losses we incur only additional logarithmic factors in $1/\varepsilon_2$. For this reason, we will restrict our attention in the below results to the logistic loss—which is convex, decreasing, 1-Lipschitz and $1/4$-smooth—although they apply equally to more general losses with different bounds that will depend on the tail behavior of the loss.

[3]This is because the first term in (2) would be bounded by $\mathsf{OPT} \cdot \sup_{|z| \le VB_X} \ell(z)$. For Lipschitz losses this incurs a term of order $O(V)$ while (for example) the exponential loss would have a term of order $O(\exp(V))$, and our proof requires $V = \Omega(\gamma^{-1})$.

We now demonstrate how to convert the bounds given in Theorem 5.2 into bounds solely involving OPT by substituting the forms of the soft margin functions given in Section 3.

Corollary 5.3 (Hard margin distributions)**.** *Suppose that $\|x\| \le B_X$ a.s. and that a unit norm population risk minimizer $\bar{v}$ for the zero-one loss satisfies $|\bar{v}^\top x| \ge \bar{\gamma} > 0$ almost surely under $\mathcal{D}_x$ for some $\bar{\gamma} > 0$. For simplicity assume that $\ell(z) = \log(1 + \exp(-z))$ is the logistic loss. Then for any $\varepsilon, \delta > 0$, with probability at least $1 - \delta$, running gradient descent for $T = \Theta(\eta^{-1}\varepsilon^{-1}\bar{\gamma}^{-2})$ with $\eta \le 2B_X^{-2}/5$ is guaranteed to find a point w_T such that*

$$\mathrm{err}_{\mathcal{D}}^{0-1}(w_T) \le \frac{1}{\log 2}\left[\mathsf{OPT} + 2B_X\bar{\gamma}^{-1}\mathsf{OPT}\log(2/\mathsf{OPT})\right] + \varepsilon,$$

provided $n = \tilde{\Omega}(\bar{\gamma}^{-2}B_X^2\log(1/\delta)\varepsilon^{-2})$.

Proof. Since $|\bar{v}^\top x| \ge \gamma^* > 0$, $\phi(\gamma^*) = 0$. Note that the logistic loss is $1/4$-smooth and satisfies $\ell^{-1}(\varepsilon) \in [\log(1/(2\varepsilon)), \log(2/\varepsilon)]$. By taking $\varepsilon_2 = \mathsf{OPT}$ the result follows by applying Theorem 5.2 with runtime $T = 4\eta^{-1}\,\varepsilon^{-1}\,\bar{\gamma}^{-2}\,\log^2(1/2\mathsf{OPT})$. □

Remark 5.4. *The bound $\tilde{O}(\bar{\gamma}^{-1}\mathsf{OPT})$ in Corollary 5.3 is tight up to logarithmic factors[4] if one wishes to use gradient descent on a convex surrogate of the form $\ell(yw^\top x)$. Diakonikolas et al. (2019, Theorem 3.1) showed that for any convex and decreasing ℓ, there exists a distribution over the unit ball with margin $\bar{\gamma} > 0$ such that a population risk minimizer $w^* := \mathrm{argmin}_w \mathbb{E}[\ell(yw^\top x)]$ has zero-one population risk at least $\Omega(\bar{\gamma}^{-1}\kappa)$, where κ is the upper bound for the Massart noise probability. The Massart noise case is more restrictive than the agnostic setting and satisfies $\mathsf{OPT} \le \kappa$. A similar matching lower bound was shown by Ben-David et al. (2012, Proposition 1).*

In the below Proposition we demonstrate the utility of having *soft* margins. As we saw in the examples in Section 3, there any many distributions that satisfy $\phi(\gamma) = O(\gamma)$. We show below the types of bounds one can expect when $\phi(\gamma) = O(\gamma^p)$ for some $p > 0$.

Proposition 5.5 (Soft margin distributions)**.** *Suppose $\|x\| \le B_X$ a.s. and that the soft margin function for a population risk minimizer of the zero-one loss satisfies $\phi(\gamma) \le C_0\gamma^p$ for some $p > 0$. For simplicity assume that ℓ is the logistic loss, and let $\eta \le (2/5)B_X^{-2}$. Assuming $\mathsf{OPT} > 0$, then for any $\varepsilon, \delta > 0$, with probability at least $1-\delta$, gradient*

[4]In fact, one can get rid of the logarithmic factors here by using the hinge loss rather than the logistic loss. In this case one needs to modify Lemma E.1 to accomodate non-smooth losses, which can be done with runtime $O(\varepsilon^{-2})$ rather than $O(\varepsilon^{-1})$ by e.g. Shalev-Shwartz & Ben-David (2014, Lemma 14.1) or using a similar argument to the one we provide for SGD in Appendix C. Then we use the fact that $\ell^{-1}(0) = 1$ for the hinge loss.

descent run for $T = \tilde{\Theta}(\eta^{-1}\varepsilon^{-1}\mathsf{OPT}^{-2/(1+p)})$ *iterations with* $n = \tilde{\Omega}(\mathsf{OPT}^{-2/(1+p)}\log(1/\delta)\varepsilon^{-2})$ *samples satisfies*

$$\mathrm{err}_{\mathcal{D}}^{0-1}(w_T) \le \tilde{O}\left((C_0 + B_X)\mathsf{OPT}^{\frac{p}{1+p}}\right) + \varepsilon,$$

Proof. By Theorem 5.2, we have

$$\mathrm{err}_{\mathcal{D}}^{0-1}(w_T) \le {}^{1}\!/_{\log 2}\Big[\big(1 + LB_X\gamma^{-1}\ell^{-1}(\varepsilon_2)\big)\,\mathsf{OPT} + C_0\gamma^p + O(\gamma^{-1}B_X\ell^{-1}(\varepsilon_2)n^{-1/2}) + \varepsilon_1 + \varepsilon_2\Big].$$

For the logistic loss, $L = 1$ and $\ell^{-1}(\varepsilon) \in [\log(1/2\varepsilon), \log(2/\varepsilon)]$ and so we take $\varepsilon_2 = \mathsf{OPT}$. Choosing $\gamma^p = \gamma^{-1}\mathsf{OPT}$, we get $\gamma = \mathsf{OPT}^{1/(1+p)}$ and hence

$$\mathrm{err}_{\mathcal{D}}^{0-1}(w_T) \le 2\big(2 + B_X\mathsf{OPT}^{-\frac{1}{1+p}}\log({}^{2}\!/_{\mathsf{OPT}})\big)\mathsf{OPT} + 2C_0\mathsf{OPT}^{\frac{1}{1+p}} + 2\varepsilon_1,$$

provided $n = \Omega(\mathsf{OPT}^{\frac{-2}{1+p}}\varepsilon_1^{-2}\log(1/\delta)\log^2({}^{1}\!/_{\mathsf{OPT}}))$ and $T = 4\eta^{-1}\varepsilon_1^{-1}\mathsf{OPT}^{-2/(1+p)}\log^2({}^{1}\!/_{2\mathsf{OPT}})$. □

By applying Proposition 5.5 to Examples 3.4 and 3.6 we get the following approximate agnostic learning guarantees for the output of gradient descent for log-concave isotropic distributions and other distributions satisfying U-anti-concentration.

Corollary 5.6. *Suppose that* $\mathcal{D}_x$ *satisfies* U*-anti-concentration and* $\|x\| \le B_X$ *a.s. Then for any* $\varepsilon, \delta > 0$*, with probability at least* $1 - \delta$*, gradient descent on the logistic loss with step size* $\eta \le (2/5)B_X^{-2}$ *and run for* $T = \tilde{O}(\eta^{-1}\varepsilon^{-1}\mathsf{OPT}^{-1})$ *iterations and* $n = \tilde{\Omega}(\mathsf{OPT}^{-1}\log(1/\delta)\varepsilon^{-2})$ *samples returns weights* w_T *satisfying* $\mathrm{err}_{\mathcal{D}}^{0-1}(w_T) \le \tilde{O}(\mathsf{OPT}^{1/2}) + \varepsilon$*, where* $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ *hide universal constant depending on* B_X*,* U*,* $\log(1/\delta)$ *and* $\log({}^{1}\!/_{\mathsf{OPT}})$ *only.*

To conclude this section, we compare our result to the variant of the Average algorithm, which estimates the vector $w_{\mathsf{Avg}} = d^{-1}\mathbb{E}_{(x,y)}[xy]$. Kalai et al. (2008) showed that when $\mathcal{D}_x$ is the uniform distribution over the unit sphere, w_{Avg} achieves risk $O(\mathsf{OPT}\sqrt{\log({}^{1}\!/_{\mathsf{OPT}})})$. Estimation of w_{Avg} can be viewed as the output convex optimization procedure, since it is the minimum of the convex objective function $F_{\mathsf{Avg}}(w) = \mathbb{E}[(\langle w, x\rangle - y)^2]$.

Although $\ell(w) = (\langle w, x\rangle - y)^2$ is convex, it is not decreasing and thus is not covered by our analysis. On the other hand, this loss function is not typically used in practice for classification problems, and the aim of this work is to characterize the guarantees for the most typical loss functions used in practice, like the logistic loss. Finally, we wish to note that the approach of soft margins is not likely to yield good bounds for the classification error when $\mathcal{D}_x$ is the uniform distribution on the unit sphere. This is because the soft margin function behavior on this distribution has a necessary dimension dependence. On the other hand, if we instead considered a scaled version of this distribution, namely $\sqrt{d}\cdot\mathrm{Unif}(\{\|x\| = 1\})$, then this dimension dependence would disappear. We provide detailed calculations for this in Appendix B.

5.2. Unbounded Distributions

We show in this section that we can achieve essentially the same results from Section 5.1 if we relax the assumption that $\mathcal{D}_x$ is bounded almost surely to being sub-exponential or possibly s-heavy-tailed.

Definition 5.7 (Sub-exponential distributions)**.** *We say* $\mathcal{D}_x$ *is* C_m-sub-exponential *if every* $x \sim \mathcal{D}_x$ *is a sub-exponential random vector with sub-exponential norm at most* C_m*. In particular, for any* $\bar{v}$ *with* $\|\bar{v}\| = 1$*,* $\mathbb{P}_{\mathcal{D}_x}(|\bar{v}^\top x| \ge t) \le \exp(-t/C_m)$.

It is well-known that log-concave isotropic distributions are C_m-sub-exponential with C_m a universal constant independent of the dimension (Balcan & Zhang, 2017).

Definition 5.8 (s-heavy tailed)**.** *We say* $\mathcal{D}_x$ *is* s-heavy-tailed *if there exists a universal constant* $C_m > 0$ *such that for every* $\bar{v} \in \mathbb{R}^d$ *with* $\|\bar{v}\| = 1$*, the probability density function* $p_{\bar{v}}(t)$ *on the subspace spanned by* $\bar{v}$ *satisfies* $p_{\bar{v}}(t) \le C_m(1 + |t|)^{-2-s}$.

We use this particular definition of s-heavy tailed so that we can more easily compare our upper bounds with the lower bounds of (Diakonikolas et al., 2020c).

As was the case for bounded distributions, the key to the proof for unbounded distributions comes from bounding the surrogate risk at a minimizer for the zero-one loss by some function of the zero-one loss.

Lemma 5.9. *Suppose* $\mathcal{D}_x$ *is* C_m*-sub-exponential. Denote by* $\bar{v}$ *as a unit norm population risk minimizer for the zero-one loss, and let* $v = V\bar{v}$ *for* $V > 0$ *be a scaled version of* $\bar{v}$*. If* ℓ *is decreasing,* L*-Lipschitz and* $\ell(0) \le 1$*, then*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(yv^\top x) \le \inf_{\gamma>0}\Big\{\phi(\gamma) + \ell(V\gamma) + \big(1 + LV + LVC_m\log({}^{1}\!/_{\mathsf{OPT}})\big)\mathsf{OPT}\Big\}.$$

If $\mathcal{D}_x$ *is only* s*-heavy tailed with constant* $C_m > 0$*, then we have*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(yv^\top x) \le \inf_{\gamma>0}\Big\{\phi(\gamma) + \ell(V\gamma) + (1 + LV)\mathsf{OPT} + LVC_m\mathsf{OPT}^{\frac{s}{1+s}}\Big\}.$$

Proof. We first show the sub-exponential case. We again use the decomposition (2), with the only difference coming

from the bound for the first term, which we show here. Fix $\xi > 0$ to be chosen later. We can write

$$\begin{aligned}
&\mathbb{E}[\ell(yv^\top x)\mathbb{1}(y\bar{v}^\top x \le 0)] \\
&\le \mathbb{E}[(1 + LV|\bar{v}^\top x|)\mathbb{1}(y\bar{v}^\top x < 0)] \\
&\quad = \mathsf{OPT} + LV\mathbb{E}[|\bar{v}^\top x|\mathbb{1}(y\bar{v}^\top x \le 0,\ |\bar{v}^\top x| \le \xi)] \\
&\qquad + LV\mathbb{E}[|\bar{v}^\top x|\mathbb{1}(y\bar{v}^\top x \le 0,\ |\bar{v}^\top x| > \xi)] \\
&\quad \le (1 + LV\xi)\mathsf{OPT} + LV\int_\xi^\infty \mathbb{P}(|\bar{v}^\top x| > t)\mathrm{d}t \qquad (8) \\
&\quad \le (1 + LV\xi)\mathsf{OPT} + LV\int_\xi^\infty \exp(-t/C_m)\mathrm{d}t \\
&\quad = (1 + LV\xi)\mathsf{OPT} + C_m LV\exp(-\xi/C_m).
\end{aligned}$$

The first inequality comes from Cauchy–Schwarz, the second from truncating, and the last from the definition of C_m-sub-exponential. Taking $\xi = C_m \log(1/\mathsf{OPT})$ results in

$$\begin{aligned}
&\mathbb{E}[\ell(yv^\top x)\mathbb{1}(y\bar{v}^\top x \le 0)] \\
&\le (1 + LV + LVC_m \log(1/\mathsf{OPT}))\,\mathsf{OPT}.
\end{aligned}$$

When $\mathcal{D}_x$ is s-heavy tailed, we can continue from (8) in a similar fashion. Denote by $p_{\bar{v}}$ as the probability density function of $\mathcal{D}_x$ on the subspace spanned by $\bar{v}$. Then we have

$$\begin{aligned}
&\mathbb{E}[\ell(yv^\top x)\mathbb{1}(y\bar{v}^\top x \le 0)] \\
&\le (1 + LV\xi)\mathsf{OPT} + LV\mathbb{E}[|\bar{v}^\top x|\mathbb{1}(|\bar{v}^\top x| > \xi)] \\
&= (1 + LV\xi)\mathsf{OPT} + LV\int_\xi^\infty t p_{\bar{v}}(t)\mathrm{d}t \\
&\le (1 + LV\xi)\mathsf{OPT} + LVC_m\int_\xi^\infty \frac{t}{(1+t)^{2+s}}\mathrm{d}t \\
&\le (1 + LV\xi)\mathsf{OPT} + LVC_m\int_\xi^\infty t^{-1-s}\mathrm{d}t \\
&= (1 + LV\xi)\mathsf{OPT} + LVC_m\xi^{-s}.
\end{aligned}$$

This bound is optimized when $\xi = \mathsf{OPT}^{\frac{-1}{1+s}}$, which leads to the desired bound. □

To derive an analogue of Theorem 5.2 for unbounded distributions, we need to extend the analysis for the generalization bound for the output of gradient descent we presented in Theorem 4.1 to unbounded distributions. Rather than using (full-batch) vanilla gradient descent, we instead use online stochastic gradient descent. The reason for this is that dealing with unbounded distributions is significantly simpler with online SGD due to the ability to work with expectations rather than high-probability bounds. It is straightforward to extend our results to vanilla gradient descent at the expense of a more involved proof by using methods from e.g., Zhang et al. (2019).

Below we present our result for unbounded distributions. Its proof is similar to that of Theorem 5.2 and can be found in Appendix C.

Theorem 5.10. *Suppose $\mathcal{D}_x$ is C_m-sub-exponential, and let $\mathbb{E}[\|x\|^2] \le B_X^2$. Let ℓ be convex, L-Lipschitz, and decreasing with $0 < \ell(0) \le 1$. Let $\varepsilon_1, \gamma > 0$ and $\varepsilon_2 \ge 0$ be arbitrary, and fix a step size $\eta \le L^{-2}B_X^{-2}\varepsilon_1/4$. By running online SGD for $T = 2\eta^{-1}\varepsilon_1^{-1}\gamma^{-2}[\ell^{-1}(\varepsilon_2)]^{-2}$ iterations after initialization at the origin, SGD finds a point w_t, $t < T$, such that in expectation over $((x_1, y_1), \dots, (x_T, y_T)) \sim \mathcal{D}^T$,*

$$\begin{aligned}
&\mathbb{E}[\mathrm{err}_{\mathcal{D}}^{0\text{-}1}(w_t)] \le {}^1\!/_{\ell(0)}\Big[\phi(\gamma) + \varepsilon_1 + \varepsilon_2 + \mathsf{OPT} \\
&+ \left(L\ell^{-1}(\varepsilon_2)\gamma^{-1} + LC_m\ell^{-1}(\varepsilon_2)\gamma^{-1}\log({}^1\!/_{\mathsf{OPT}})\right)\mathsf{OPT}\Big].
\end{aligned}$$

If instead we only know $\mathcal{D}_x$ is s-heavy tailed, we have

$$\begin{aligned}
&\mathbb{E}[\mathrm{err}_{\mathcal{D}}^{0\text{-}1}(w_t)] \le {}^1\!/_{\ell(0)}\Big[\phi(\gamma) + \varepsilon_1 + \varepsilon_2 \\
&+ (1 + L\ell^{-1}(\varepsilon_2)\gamma^{-1})\mathsf{OPT} + LC_m\ell^{-1}(\varepsilon_2)\gamma^{-1}\mathsf{OPT}^{\frac{s}{1+s}}\Big].
\end{aligned}$$

The above theorem yields the following bound for sub-exponential distributions and heavy-tailed distributions satisfying U-anti-concentration. Recall from Example 3.6 that log-concave isotropic distributions are $O(1)$-sub-exponential and satisfy anti-concentration with $U = 1$.

Corollary 5.11. *Suppose $\mathcal{D}_x$ is C_m-sub-exponential with $\mathbb{E}[\|x\|^2] \le B_X^2$ and assume U-anti-concentration holds. Let ℓ be the logistic loss and let $\varepsilon > 0$. Fix a step size $\eta \le B_X^{-2}\varepsilon/16$. By running online SGD for $T = \tilde{O}(\eta^{-1}\varepsilon^{-1}C_m U^{-1}\mathsf{OPT}^{-1})$ iterations, there exists a point w_t, $t < T$, such that*

$$\mathbb{E}[\mathrm{err}_{\mathcal{D}}^{0\text{-}1}(w_t)] \le \tilde{O}\left(\mathsf{OPT}^{1/2}\right) + \varepsilon.$$

If instead we only know $\mathcal{D}_x$ is s-heavy tailed, we have

$$\mathbb{E}[\mathrm{err}_{\mathcal{D}}^{0\text{-}1}(w_t)] \le \tilde{O}\left(\mathsf{OPT}^{\frac{s}{2(1+s)}}\right) + \varepsilon.$$

Proof. Consider sub-exponential distributions first. By Example 3.4, $\phi(\gamma) \le 2\gamma U$. Since $\ell^{-1}(\varepsilon) \in [\log(1/2\varepsilon), \log(2/\varepsilon)]$, we can take $\varepsilon_2 = \mathsf{OPT}$ in Theorem 5.10 to get

$$\begin{aligned}
&\mathbb{E}[\mathrm{err}_{\mathcal{D}}^{0\text{-}1}(w_t)] \le {}^1\!/_{\log(2)}\Big[2\gamma U + \varepsilon \\
&\quad + \left(2 + C_m + LC_m\gamma^{-1}\log^2({}^2\!/_{\mathsf{OPT}})\right)\mathsf{OPT}\Big].
\end{aligned}$$

This bound is optimized when $U\gamma = C_m\gamma^{-1}\mathsf{OPT}$, i.e., $\gamma = U^{-1/2}C_m^{1/2}\mathsf{OPT}^{\frac{1}{2}}$. Substituting this value for γ we get the desired bound with $T = 2\log(2)\eta^{-1}\varepsilon^{-1}C_m U^{-1}\mathsf{OPT}^{-1}\log^2({}^1\!/_{2\mathsf{OPT}})$.

For s-heavy tailed distributions, the argument is essentially the same. We again use $\phi(\gamma) \le 2\gamma U$ and take $\varepsilon_2 = \mathsf{OPT}$. The optimal choice of γ occurs when $\gamma^{-1}\mathsf{OPT}^{\frac{s}{1+s}} = \gamma$. Solving for γ gives $\gamma = \mathsf{OPT}^{\frac{s}{2(1+s)}}$ and the result follows by substituting γ into Theorem 5.10. □

Remark 5.12. *Diakonikolas et al. (2020d, Theorem 1.4) recently showed that if the marginal of $\mathcal{D}$ over x is the standard Gaussian in d dimensions, for every convex, non-decreasing loss ℓ, the minimizer $v = \operatorname{argmin}_w F_\ell(w)$ satisfies $\operatorname{err}_{\mathcal{D}}^{0-1}(v) = \Omega(\mathsf{OPT}\sqrt{\log(1/\mathsf{OPT})})$, in comparison with our upper bound of $\tilde{O}(\mathsf{OPT}^{1/2})$. For s-heavy-tailed distributions, their lower bound is $\operatorname{err}_{\mathcal{D}}^{0-1}(v) = \Omega(\mathsf{OPT}^{1-\frac{1}{s}})$, so that as $s \to 2$ our upper bound tends to $\tilde{O}(\mathsf{OPT}^{1/3})$ in comparison to their lower bound of $\Omega(\mathsf{OPT}^{1/2})$. Further narrowing the gap between our upper bounds and their lower bounds is an interesting open problem.*

We also wish to note that Diakonikolas et al. (2020d) showed that by using gradient descent on a certain bounded and decreasing non-convex surrogate for the zero-one loss, it is possible to show that gradient descent finds a point with $\operatorname{err}_{\mathcal{D}}^{0-1}(w_T) \leq O(\mathsf{OPT}) + \varepsilon$. In comparison with our result, this is perhaps not surprising: if one is able to show that gradient descent with a bounded *and decreasing loss function can achieve population risk bounded by $O(\mathbb{E}[\ell(yv^\top x)])$ for arbitrary $v \in \mathbb{R}^d$, then the same proof technique that yields Theorem 5.10 from Lemma 5.9 would demonstrate that $\operatorname{err}_{\mathcal{D}}^{0-1}(w_t) \leq O(\mathsf{OPT})$. Since the only globally bounded convex function is constant, this approach would require working with a non-convex loss.*

6. Conclusion and Future Work

In this work we analyzed the problem of learning halfspaces in the presence of agnostic label noise. We showed that the simple approach of gradient descent on convex surrogates for the zero-one loss (such as the cross entropy or hinge losses) can yield approximate minimizers for the zero-one loss for both hard margin distributions and sub-exponential distributions satisfying an anti-concentration inequality enjoyed by log-concave isotropic distributions. Our approach relied upon developing a novel connection between minimizers of convex surrogates of the zero-one loss to minimizers of the zero-one loss itself, with the soft margin property playing a key role in this connection. Our results match (up to logarithmic factors) lower bounds shown for hard margin distributions. For future work, we are interested in exploring the utility of the soft margin for understanding other classification problems.

Acknowledgements

We thank Peter Bartlett for helpful comments that led us to the result on fast rates for stochastic gradient descent. We also thank the anonymous reviewers for their helpful comments. SF is supported by the UCLA Dissertation Year Fellowship. YC and QG are partially supported by the National Science Foundation IIS-2008981. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Angluin, D. and Laird, P. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Awasthi, P., Balcan, M.-F., Haghtalab, N., and Urner, R. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory (COLT)*, 2015.

Awasthi, P., Balcan, M.-F., Haghtalab, N., and Zhang, H. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory (COLT)*, 2016.

Awasthi, P., Balcan, M., and Long, P. M. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6):50:1–50:27, 2017.

Balcan, M.-F. and Haghtalab, N. Noise in classification. In Roughgarden, T. (ed.), *Beyond Worst Case Analysis of Algorithms*, chapter 16. Cambridge University Press, 2021.

Balcan, M.-F. F. and Zhang, H. Sample and computationally efficient learning algorithms under s-concave distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. (Was Department of Statistics, U.C. Berkeley Technical Report number 638, 2003).

Ben-David, S., Loker, D., Srebro, N., and Sridharan, K. Minimizing the misclassification error rate using a surrogate convex loss. In *International Conference on Machine Learning (ICML)*, 2012.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandit algorithms with supervised learning guarantees. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Blum, A., Frieze, A., Kannan, R., and Vempala, S. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1-2):35–52, 1998.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Conference on Learning Theory (COLT)*, 1992.

Cao, Y. and Gu, Q. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

Daniely, A. Complexity theoretic limitations on learning halfspaces. In *ACM Symposium on Theory of Computing (STOC)*, pp. 105–117, 2016.

Diakonikolas, I., Gouleakis, T., and Tzamos, C. Distribution-independent pac learning of halfspaces with massart noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Diakonikolas, I., Goel, S., Karmalkar, S., Klivans, A. R., and Soltanolkotabi, M. Approximation schemes for relu regression. In *Conference on Learning Theory (COLT)*, 2020a.

Diakonikolas, I., Kane, D. M., and Zarifis, N. Near-optimal sq lower bounds for agnostically learning halfspaces and relus under gaussian marginals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory (COLT)*, 2020c.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Non-convex sgd learns halfspaces with adversarial label noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020d.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning halfspaces with tsybakov noise. *arXiv preprint arXiv:2006.06467*, 2020e.

Foster, D. J., Sekhari, A., and Sridharan, K. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, 2018.

Frei, S., Cao, Y., and Gu, Q. Algorithm-dependent generalization bounds for overparameterized deep residual networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Frei, S., Cao, Y., and Gu, Q. Agnostic learning of a single neuron with gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Goel, S., Karmalkar, S., and Klivans, A. R. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems 32*, 2019.

Goel, S., Gollakota, A., and Klivans, A. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Guruswami, V. and Raghavendra, P. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39 (2):742–765, 2009.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Ji, Z., Dudík, M., Schapire, R. E., and Telgarsky, M. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory (COLT)*, 2020.

Kalai, A. T., Klivans, A. R., Mansour, Y., and Servedio, R. A. Agnostically learning halfspaces. *SIAM J. Comput.*, 37(6):1777–1805, 2008.

Kearns, M. J., Schapire, R. E., and Sellie, L. M. Toward efficient agnostic learning. *Machine Learning*, 17(2-3): 115–141, 1994.

Massart, P., Nédélec, É., et al. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Servedio, R. A. On pac learning using winnow, perceptron, and a perceptron-like algorithm. In *Conference on Computational Learning Theory*, pp. 296–307, 1999.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

Shamir, O. Gradient methods never overfit on separable data. *arXiv preprint*, arXiv:2007.00028, 2020.

Sloan, R. Types of noise in data for concept learning. In *Conference on Learning Theory (COLT)*, 1988.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

Sridharan, K., Shalev-Shwartz, S., and Srebro, N. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009.

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

Zhang, X., Yu, Y., Wang, L., and Gu, Q. Learning one-hidden-layer relu networks via gradient descent. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

A. Fast Rates with Stochastic Gradient Descent

In Theorem 4.1, we showed that $F_\ell(w_T) \leq F_\ell(v) + O(1/\sqrt{n})$ given n samples from $\mathcal{D}$ by using vanilla (full-batch) gradient descent. In this section we demonstrate that by instead using stochastic gradient descent, one can achieve $F_\ell(w_T) \leq O(F_\ell(v)) + O(1/n)$ by appealing to a martingle Bernstein bound. We note that although the population risk guarantee degrades from $F_\ell(v)$ to $O(F_\ell(v))$, our bounds for the zero-one risk in vanilla gradient descent already have constant-factor errors and so the constant-factor error for $F_\ell(v)$ will not change the order of our final bounds.

The version of stochastic gradient descent that we study is the standard online SGD. Suppose we sample $z_t = (x_t, y_t) \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ for $t = 1, \ldots, T$, and let us denote the σ-algebra generated by the first t samples as $\mathcal{G}_t = \sigma(z_1, \ldots, z_t)$. Define

$$\widehat{F}_t(w) := \ell(y_t w^\top x_t), \quad \mathbb{E}[\widehat{F}_t(w_t)|\mathcal{G}_{t-1}] = F(w_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(y w_t^\top x).$$

The online stochastic gradient descent updates take the form

$$w_{t+1} := w_t - \eta \nabla \widehat{F}_t(w_t).$$

We are able to show an improved rate of $O(\varepsilon^{-1})$ when using online SGD.

Theorem A.1 (Fast rate for online SGD)**.** *Assume that $\ell(\cdot) \geq 0$ is convex, strictly decreasing, L-Lipschitz and H-smooth. Assume $\|x\| \leq B_X$ a.s. For simplicity assume that $w_0 = 0$. Let $v \in \mathbb{R}^d$ be arbitrary with $\|v\| \leq V$. Let $\eta \leq (32HB_X^2)^{-1}$. Then for any $\varepsilon, \delta > 0$, by running online stochastic gradient descent for $T = O(\varepsilon^{-1}V^2\log(1/\delta))$ iterations, with probability at least $1-\delta$ there exists a point w_{t^*}, with $t^* < T$, such that*

$$\mathrm{err}_{\mathcal{D}}^{0-1}(w_{t^*}) \leq O(\mathbb{E}[\ell(yv^\top x)]) + \varepsilon,$$

where $O(\cdot)$ hides constant factors that depend on L, H and B_X only.

In this section we will sketch the proof for the above theorem. First, we note the following guarantee for the empirical risk. This result is a standard result in online convex optimization (see, e.g., Theorem 14.13 in Shalev-Shwartz & Ben-David (2014)).

Lemma A.2. *Suppose that $\ell(\cdot) \geq 0$ is convex and H-smooth, and that $\|x\| \leq B_X$ a.s. Then for any $\alpha \in (0,1)$, for fixed step size $\eta \leq \alpha/(8HB_X^2)$, and for any $T \geq 1$, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\widehat{F}_t(w_t) \leq (1+\alpha)\frac{1}{T}\sum_{t=0}^{T-1}\widehat{F}_t(v) + \frac{\|w_0 - v\|^2}{\eta T}.$$

From here, one could take expectations and show that in expectation over the randomness of SGD, the population risk found by gradient descent is at most $(1+\alpha)F_\ell(v) + O(1/T)$, but we are interested in developing a generalization bound that has the same fast rate but holds with high probability, which requires significantly more work. Much of the literature for fast rates in stochastic optimization require additional structure to achieve such results: Bartlett et al. (2006) showed that the empirical risk minimizer converges at a fast rate to its expectation under a low-noise assumption; Sridharan et al. (2009) achieved fast rates for the output of stochastic optimization by using explicit regularization by a strongly convex regularizer; Srebro et al. (2010) shows that projected online SGD achieves fast rates when $\min_v \mathbb{E}[\ell(yv^\top x)] = 0$. By contrast, we show below that the standard online SGD algorithm achieves a constant-factor approximation to the best population risk at a fast rate. We do so by appealing to the following martingale Bernstein inequality.

Lemma A.3 (Beygelzimer et al. (2011), Theorem 1)**.** *Let $\{Y_t\}$ be a martingale adapted to the filtration $\mathcal{G}_t$, and let $Y_0 = 0$. Let $\{D_t\}$ be the corresponding martingale difference sequence. Fix $T > 0$, and define the sequence of conditional variance*

$$U_{T-1} := \sum_{t<T}\mathbb{E}[D_t^2|\mathcal{G}_{t-1}],$$

and assume that $D_t \leq R$ almost surely. Then for any $\delta \in (0,1)$, with probability greater than $1-\delta$,

$$Y_{T-1} \leq R\log(1/\delta) + (e-2)U_{T-1}/R.$$

We would like to take $Y_t = \sum_{\tau<t}[F(w_t) - \widehat{F}_t(w_t)]$, which has martingale difference sequence $D_t = F(w_t) - \widehat{F}_t(w_t)$. The difficulty here is showing that $D_t \leq R$ almost surely for some absolute constant R. The obvious fix would be to show that the weights w_t stay within a bounded region throughout gradient descent via early stopping. In the case of full-batch gradient descent, this is indeed possible: in Lemma E.1 we showed that $\|w_t - v\| \leq \|w_0 - v\|$ throughout gradient descent, which would imply that $\ell(yw_t^\top x)$ is uniformly bounded for all samples x throughout G.D., in which case $D_t \leq F(w_t)$ would hold almost surely. But for online stochastic gradient descent, since we must continue to take draws from the distribution in order to reduce the optimization error, there isn't a straightforward way to get a bound on $\|w_t\|$ to hold almost surely throughout the gradient descent trajectory.

Our way around this is to realize that in the end, our end goal is to show something of the form

$$\mathrm{err}_{\mathcal{D}}^{0-1}(w_t) \leq O(\mathbb{E}[\ell(yv^\top x)]) + O(1/T),$$

since then we could use a decomposition similar to Lemma 5.1 to bound the right hand side by terms involving OPT and a soft margin function. Since for a non-negative H-smooth loss $[\ell'(z)]^2 \leq 4H\ell(z)$ holds, it actually suffices to show that the losses $\{[\ell'(y_t w^\top x_t)]^2\}_1^T$ concentrate around their expectation at a fast rate. Roughly, this is because one would have

$$\begin{aligned}
\min_{t<T} \mathbb{E}_{\mathcal{D}}\left([\ell'(yw_t^\top x)]^2\right) &\leq \frac{1}{T}\sum_{t=0}^{T-1}[\ell'(y_t w_t^\top x_t)^2] + O(1/T) \\
&\leq \frac{4H}{T}\sum_{t=0}^{T-1}\ell(y_t w_t^\top x_t) + O(1/T) \\
&\leq \frac{4H}{T}\sum_{t=0}^{T-1}\ell(y_t v^\top x_t) + O(1/T). \qquad (9)
\end{aligned}$$

To finish the proof we can then use the fact that v is a fixed vector of constant norm to show that the empirical risk on the last line of (9) concentrates around $O(\mathbb{E}[\ell(yv^\top x)])$ at rate $O(1/T)$. For decreasing and convex loss functions, $\ell'(z)^2$ is decreasing so the above provides a bound for $\mathrm{err}_{\mathcal{D}}^{0-1}(w_t)$ by Markov's inequality.

This shows that the key to the proof is to show that $\{\ell'(y_t w_t x_t)^2\}$ concentrates at rate $O(1/T)$. The reason this is easier than showing concentration of $\{\ell(y_t w_t x_t)\}$ is because for Lipschitz losses, $\ell'(y_t w_t^\top x_t)^2$ is uniformly bounded regardless of the norm of w_t. This ensures that the almost sure condition needed for the martingale difference sequence in Lemma A.3 holds trivially. We note that a similar technique has been utilized before for the analysis of SGD (Ji & Telgarsky, 2020; Cao & Gu, 2020; Frei et al., 2019), although in these settings the authors used the concentration of $\{\ell'(z_t)\}$ rather than $\{\ell'(z_t)^2\}$ since they considered the logistic loss, for which $|\ell'(z)| \leq \ell(z)$. Since not all smooth loss functions satisfy this inequality, we instead use concentration of $\{\ell'(z_t)^2\}$.

Below we formalize the above proof sketch. We first show that $\{\ell'(y_t w_t^\top x_t)^2\}$ concentrates at rate $O(1/T)$ for any fixed sequence of gradient descent iterates $\{w_t\}$.

Lemma A.4. *Let ℓ be any differentiable L-Lipschitz function, and let $z_t = (x_t, y_t) \overset{\text{i.i.d.}}{\sim} \mathcal{D}$. Denote $\mathcal{G}_t = \sigma(z_1, \ldots, z_t)$ the σ-algebra generated by the first t draws from $\mathcal{D}$, and let $\{w_t\}$ be any sequence of random variables such that w_t is $\mathcal{G}_{t-1}$-measurable for each t. Then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left(\left[\ell'(yw_t^\top x)\right]^2\right) \leq \frac{4}{T}\sum_{t=0}^{T-1}\left[\ell'(y_t w_t^\top x_t)\right]^2 + \frac{4L^2\log(1/\delta)}{T}. \qquad (10)$$

Proof. For simplicity, let us denote

$$J(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left(\left[\ell'(yw^\top x)\right]^2\right), \quad \widehat{J}_t(w) := \left[\ell'(y_t w^\top x_t)\right]^2.$$

We begin by showing the second inequality in (10). Define the random variable

$$Y_t := \sum_{\tau<t}(J(w_\tau) - \widehat{J}_\tau(w_\tau)) \qquad (11)$$

Then Y_t is a martingale with respect to the filtration $\mathcal{G}_{t-1}$ with martingale difference sequence $D_t := J(w_t) - \widehat{J}_t(w_t)$. We need bounds on D_t and on $\mathbb{E}[D_t^2|\mathcal{G}_{t-1}]$ in order to apply Lemma A.3. Since ℓ is L-Lipschitz,

$$D_t \leq J(w_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left([-\ell'(yv^\top x)]^2\right) \leq L^2.$$

Similarly,

$$\begin{aligned}\mathbb{E}[\widehat{J}_t(w_t)^2|\mathcal{G}_{t-1}] &= \mathbb{E}\left(\left[\ell'(y_t w_t^\top x_t)\right]^4|\mathcal{G}_{t-1}\right)\\ &\leq L^2\mathbb{E}\left(\left[\ell'(y_t w_t^\top x_t)\right]^2|\mathcal{G}_{t-1}\right)\\ &= L^2 J(w_t). \end{aligned} \tag{12}$$

In the inequality we use that ℓ is L-Lipschitz, so that $|\ell'(\alpha)| \leq L$. We then can use (12) to bound the squared increments,

$$\begin{aligned}\mathbb{E}[D_t^2|\mathcal{G}_{t-1}] &= J(w_t)^2 - 2J(w_t)\mathbb{E}[\widehat{J}_t(w_t)|\mathcal{G}_{t-1}] + \mathbb{E}[\widehat{J}_t(w_t)^2|\mathcal{G}_{t-1}]\\ &\leq \mathbb{E}[\widehat{J}_t(w_t)^2|\mathcal{G}_{t-1}]\\ &\leq L^2 J(w_t).\end{aligned}$$

This allows for us to bound

$$U_{T-1} = \sum_{t=0}^{T-1}\mathbb{E}[D_t^2|\mathcal{G}_{t-1}] \leq L^2\sum_{t=0}^{T-1} J(w_t).$$

Lemma A.3 thus implies that with probability at least $1-\delta$, we have

$$\sum_{t=0}^{T-1}(J(w_t) - \widehat{J}_t(w_t)) \leq L^2\log(1/\delta) + (\exp(1)-2)\sum_{t=0}^{T-1} J(w_t).$$

Using that $(1-\exp(1)+2)^{-1} \leq 4$, we divide each side by T and get

$$\frac{1}{T}\sum_{0=t}^{T-1} J(w_t) \leq \frac{4}{T}\sum_{t=0}^{T-1}\widehat{J}_t(w_t) + \frac{4L^2\log(1/\delta)}{T}.$$

This completes the proof. □

Next, we show that the average of $\{\ell(y_t v^\top x_t)\}$ is at most twice its mean at rate $O(1/T)$.

Lemma A.5. *Let ℓ be any L-Lipschitz function, and suppose that $\ell(0) \leq 1$ and $\|x\|_2 \leq B$ a.s. Let $v \in \mathbb{R}^d$ be arbitrary with $\|v\| \leq V$. For any $\delta > 0$, with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\widehat{F}_t(v) \leq 2F(v) + \frac{2(1+LVB_X)\log(1/\delta)}{T}.$$

Proof. Let $\mathcal{G}_t = \sigma(z_1,\ldots,z_t)$ be the σ-algebra generated by the first t draws from $\mathcal{D}$. Then the random variable $Y_t := \sum_{\tau<t}(\widehat{F}_\tau(v) - F(v))$ is a martingale with respect to the filtration $\mathcal{G}_{t-1}$ with martingale difference sequence $D_t := \widehat{F}_t(v) - F(v)$. We need bounds on D_t and on $\mathbb{E}[D_t^2|\mathcal{G}_{t-1}]$ in order to apply Lemma A.3. Since ℓ is L-Lipschitz and $\|x\| \leq B_X$ a.s., that $\|v\| \leq V$ implies that almost surely,

$$D_t \leq \widehat{F}_t(v) = \ell(y_t v^\top x_t) \leq (1 + LVB_X). \tag{13}$$

Similarly,

$$\begin{aligned}\mathbb{E}[\widehat{F}_t(v)^2|\mathcal{G}_{t-1}] &= \mathbb{E}\left[\ell(y_t v^\top x_t)^2|\mathcal{G}_{t-1}\right]\\ &\leq (1+LVB_X)\mathbb{E}[\ell(y_t v^\top x_t)]\\ &= (1+LVB_X)F(v).\end{aligned} \tag{14}$$

In the inequality, we have used that (x_t, y_t) is independent from $\mathcal{G}_{t-1}$ together with (13). We then can use (14) to bound the squared increments,

$$\begin{aligned}\mathbb{E}[D_t^2|\mathcal{G}_{t-1}] &= F(v)^2 - 2F(v)\mathbb{E}[\widehat{F}_t(v)|\mathcal{G}_{t-1}] + \mathbb{E}[\widehat{F}_t(v)^2|\mathcal{G}_{t-1}] \\ &\leq \mathbb{E}[\widehat{F}_t(v)^2|\mathcal{G}_{t-1}] \\ &\leq (1 + LVB_X)F(v).\end{aligned}$$

This allows for us to bound

$$U_{T-1} := \sum_{t=0}^{T-1} \mathbb{E}[D_t^2|\mathcal{G}_{t-1}] \leq (1 + LVB_X)TF(v).$$

Lemma A.3 thus implies that with probability at least $1 - \delta$, we have

$$\sum_{t=0}^{T-1}(\widehat{F}_t(v) - F(v)) \leq (1 + LVB_X)\log(1/\delta) + (\exp(1) - 2)TF(v).$$

Using that $\exp(1) - 2 \leq 1$, we divide each side by T and get

$$\frac{1}{T}\sum_{t=0}^{T-1}\widehat{F}_t(v) \leq 2F(v) + \frac{2(1 + LVB_X)\log(1/\delta)}{T}.$$

□

Finally, we put these ingredients together for the proof of Theorem A.1.

Proof. Since ℓ is convex and H-smooth, we can take $\alpha = 1/4$ in Lemma A.2 to get

$$\frac{1}{T}\sum_{t=0}^{T-1}\widehat{F}_t(w_t) \leq \frac{5}{4T}\sum_{t=0}^{T-1}\widehat{F}_t(v) + \frac{V^2}{\eta T}. \tag{15}$$

We can therefore bound

$$\begin{aligned}\min_{t<T}\mathbb{E}\left([\ell'(yw_t^\top x)]^2\right) &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left([\ell'(yw_t^\top x)]^2\right) \\ &\leq \frac{4}{T}\sum_{t=0}^{T-1}[\ell'(y_t w_t^\top x_t)]^2 + \frac{4L^2\log(2/\delta)}{T} \\ &\leq \frac{16H}{T}\sum_{t=0}^{T-1}\widehat{F}_t(w_t) + \frac{4L^2\log(2/\delta)}{T} \\ &\leq \frac{20H}{T}\sum_{t=0}^{T-1}\widehat{F}_t(v) + \frac{5L^2\log(2/\delta) + V^2}{\eta T} \\ &\leq 40HF(v) + \frac{40H(1 + LVB_X)\eta\log(2/\delta) + 5L^2\eta\log(2/\delta) + V^2}{\eta T}.\end{aligned} \tag{16}$$

The second inequality holds since ℓ is L-Lipschitz so that we can apply Lemma A.4. The third inequality uses that ℓ is non-negative and H-smooth, so that $[\ell'(z)]^2 \leq 4H\ell(z)$ (see Srebro et al. (2010, Lemma 2.1)). The fourth inequality uses (15), and the final inequality uses Lemma A.5.

Since ℓ is convex and decreasing, $\frac{\mathrm{d}}{\mathrm{d}z}\ell'(z)^2 = 2\ell'(z)\ell''(z) \leq 0$, so $\ell'(z)^2$ is decreasing. By Markov's inequality, this implies

$$\mathbb{P}(yw_t^\top x < 0) = \mathbb{P}\left([\ell'(yw_t^\top x)]^2 \geq \ell'(0)^2\right) \leq [\ell'(0)]^{-2}\mathbb{E}\left([\ell'(yw_t^\top x)]^2\right).$$

Substituting this into (16), this implies that with probability at least $1 - \delta$,

$$\mathrm{err}_{\mathcal{D}}^{0-1}(w_t) \leq O(F(v)) + O(V^2\log(1/\delta)/T).$$

□

We note that the above proof works for an arbitrary initialization w_0 such that $\|w_0\|$ is bounded by an absolute constant with high probability, e.g. with the random initialization $w_0 \overset{\text{i.i.d.}}{\sim} N(0, I_d/d)$. The only difference is that we need to replace V^2 with $\|w_0 - v\|^2 \le O(V^2)$ in (15) and the subsequent lines.

B. Soft Margin for Uniform Distribution

We show here that the soft margin function for the uniform distribution on the sphere has an unavoidable dimension dependence. Consider $x \sim \mathcal{D}$ is uniform on the sphere in d dimensions. Then x has the same distribution as $z/\|z\|$, where $z \sim N(0, I_d)$ is the d-dimensional Gaussian. The soft margin function on x thus satisfies, for $\|v\| = 1$,

$$\phi(\gamma) = \mathbb{P}_x(|v^\top x| \le \gamma) = \mathbb{P}_z\left(|v^\top z|^2/\|z\|^2 \le \gamma^2\right).$$

By symmetry, we can rotate the coordinate system so that $v = (1, 0, \dots)$, which results in $\phi(\gamma)$ taking the form

$$\begin{aligned}\mathbb{P}\left(\frac{z_1^2}{\sum_{i=1}^d z_i^2} \le \gamma^2\right) &= \mathbb{P}\left((1-\gamma^2)z_1^2 \le \gamma^2 \textstyle\sum_{i=2}^d z_i^2\right) \\ &= \mathbb{P}\left(z_1^2 \le \frac{\gamma^2}{1-\gamma^2}\textstyle\sum_{i=2}^d z_i^2\right) \ge \mathbb{P}(z_1^2 \le \gamma^2 \textstyle\sum_{i=2}^d z_i^2).\end{aligned}$$

Since $\gamma^2 \sum_{i=2}^d z_i^2 = \Theta(\gamma^2 d)$ with high probability by concentration of the χ^2 distribution, and since $\mathbb{P}(|z_1| \le a) = \Theta(a)$ for the Gaussian, this shows that $\phi(\gamma) = \Omega(\gamma\sqrt{d})$ when $\mathcal{D}_x$ is uniform on the sphere. Thus our approach of using the soft margin in Theorem 5.2 to derive generalization bounds will result in multiplicative terms attached to OPT that will grow with d for such a distribution.

On the other hand, note that if we instead consider the uniform distribution on the sphere scaled by $\sqrt{d}$, then the same argument about would yield

$$\begin{aligned}\phi(\gamma) &= \mathbb{P}\left(d\frac{z_1^2}{\sum_{i=1}^d z_i^2} \le \gamma^2\right) \\ &= \mathbb{P}\left(z_1^2 \le \frac{\gamma^2}{1-\gamma^2} \cdot \frac{1}{d}\sum_{i=2}^d z_i^2\right) \\ &= \mathbb{P}\left(|z_1| \le \sqrt{\frac{\gamma^2}{1-\gamma^2}} \cdot \sqrt{\frac{1}{d}\sum_{i=2}^d z_i^2}\right)\end{aligned}$$

By concentration of the $\chi^2(d)$ distribution, $\frac{1}{d}\sum_{i=2}^d z_i^2 = \Theta((d-1)/d) = \Theta(1)$ w.h.p. and thus we could use the 1-anti-concentration of the standard d dimensional Gaussian to get $\phi(\gamma) \le O(\sqrt{\gamma^2/(1-\gamma^2)}) = O(\gamma)$.

C. Proofs for Unbounded Distributions

In this section we prove Theorem 5.10.

C.1. Empirical Risk

First, we derive an analogue of Lemma E.1 that holds for any distribution satisfying $\mathbb{E}[\|x\|^2] \le B_X^2$ by appealing to online stochastic gradient descent. Note that any distribution over $\mathbb{R}^d$ with sub-Gaussian coordinates satisfies $\mathbb{E}[\|x\|^2] \le B^2$ for some $B \in \mathbb{R}$.

We use the same notation from Section A, where we assume samples $z_t = (x_t, y_t) \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ for $t = 1, \dots, T$, and $\mathcal{G}_t := \sigma(z_1, \dots, z_t)$, and denote

$$\widehat{F}_t(w) := \ell(y_t w^\top x_t), \quad \mathbb{E}[\widehat{F}_t(w_t)|\mathcal{G}_{t-1}] = F(w_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(y w_t^\top x).$$

The online stochastic gradient descent updates take the form

$$w_{t+1} := w_t - \eta \nabla \widehat{F}_t(w_t).$$

Lemma C.1. *Suppose $\mathbb{E}_{\mathcal{D}_x}[\|x\|^2] \le B_X^2$. Suppose that ℓ is convex and L-Lipschitz. Let $v \in \mathbb{R}^d$ and $\varepsilon, \alpha \in (0,1)$ be arbitrary, and consider any initialization $w_0 \in \mathbb{R}^d$. Provided $\eta \le L^{-2} B_X^{-2} \varepsilon/2$, then for any $T \in \mathbb{N}$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E} F(w_t) \le F(v) + \frac{\|w_0 - v\|^2}{\eta T} + \varepsilon.$$

Proof. The proof is very similar to that of the proof of Lemma A.2 described in Appendix C.1, so we describe here the main modifications. The key difference comes from the gradient upper bound: for $g_t = \ell'(y_t w_t^\top x_t)$, instead of getting an upper bound that holds a.s. in terms of the loss, we only show that its expectation is bounded by a constant:

$$\mathbb{E}[\|g_t\|^2 \,|\mathcal{G}_{t-1}] \le \mathbb{E}[\ell'(y_t w_t x_t)^2 \|x_t\|^2 \,|\mathcal{G}_{t-1}] \le L^2 \mathbb{E}[\|x_t\|^2 \,|\mathcal{G}_{t-1}] \le L^2 B_X^2.$$

By convexity, $\langle g_t, w_t - v\rangle \ge \widehat{F}_t(w_t) - \widehat{F}_t(v)$. Thus taking $\eta = O(\varepsilon)$, we get

$$\begin{aligned}\|w_t - v\|^2 - \mathbb{E}[\|w_{t+1} - v\|^2 \,|\mathcal{G}_{t-1}] &\ge \mathbb{E}[2\eta(\widehat{F}_t(w_t) - \widehat{F}_t(v)) - \eta^2 \|g_t\|^2 \,|\mathcal{G}_{t-1}] \\ &\ge 2\eta(F(w_t) - F(v)) - \eta^2 L^2 B_X^2 \\ &\ge 2\eta(F(w_t) - F(v) - \varepsilon).\end{aligned}$$

Taking expectations with respect to the randomness of SGD and summing from 0 to $T-1$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E} F(w_t) \le F(v) + \frac{\|w_0 - v\|^2}{\eta T} + \varepsilon.$$

□

We note that the above analysis is quite loose and we are aware of a number of ways to achieve faster rates by introducing various assumptions on ℓ and $\mathcal{D}_x$; we chose the presentation above for simplicity.

With the above result in hand, we can prove Theorem 5.10.

Proof. Consider sub-exponential distributions first. Let $\varepsilon_1 > 0$. By taking $\eta \le L^{-2} B_X^{-2} \varepsilon_1/8$ and $T = 2V^2 \eta^{-1} \varepsilon_1^{-1}$, Lemma C.1 and Markov's inequality, this implies that there exists some $t < T$ such that

$$\mathbb{E}[\mathrm{err}_{\mathcal{D}}^{0-1}(w_t)] \le \mathbb{E}[F(w_t)] \le {}^1\!/_{\ell(0)} \Big[F(v) + \frac{V^2}{\eta T} + \varepsilon_1/2 \le F(v) + \varepsilon_1\Big].$$

By Lemma 5.9, this implies that for any $\gamma > 0$,

$$\mathbb{E}[\mathrm{err}_{\mathcal{D}}^{0-1}(w_t)] \le {}^1\!/_{\ell(0)} \Big[(1 + C_m + LVC_m \log({}^1\!/_{\mathsf{OPT}}))\, \mathsf{OPT} + \phi(\gamma) + \ell(V\gamma) + \varepsilon_1\Big].$$

For $\varepsilon_2 \ge 0$, by taking $V = \gamma^{-1}\ell^{-1}(\varepsilon_2)$, this means that for any $\gamma > 0$, we have

$$\mathbb{E}[\mathrm{err}_{\mathcal{D}}^{0-1}(w_t)] \le {}^1\!/_{\ell(0)} \Big[\big(1 + C_m + LC_m \ell^{-1}(\varepsilon_2)\gamma^{-1} \log({}^1\!/_{\mathsf{OPT}})\big)\, \mathsf{OPT} + \phi(\gamma) + \varepsilon_1 + \varepsilon_2\Big].$$

For $V = \gamma^{-1}\ell^{-1}(\varepsilon_2)$, we need $T = 2\gamma^{-2}\eta^{-1}\varepsilon_1^{-1}[\ell^{-1}(\varepsilon_2)]^2$.

For heavy-tailed distributions, the proof is essentially identical except now we use the heavy-tailed part of Lemma 5.9:

$$\mathbb{E}[\mathrm{err}_{\mathcal{D}}^{0-1}(w_t)] \le {}^1\!/_{\ell(0)} \Big[\phi(\gamma) + \varepsilon_1 + \varepsilon_2 + (1 + L\ell^{-1}(\varepsilon_2)\gamma^{-1})\mathsf{OPT} + LC_m \ell^{-1}(\varepsilon_2)\gamma^{-1}\mathsf{OPT}^{\frac{s}{1+s}}\Big].$$

The same choice of V and T gives the result.

□

D. Loss Functions and Sample Complexity for Separable Data

We present here the proof of Corollary 4.2.

Proof. Let $v = V\bar{v}$. By Theorem 4.1, for any $\varepsilon, \delta > 0$ and $V > 0$, running gradient descent for $T = 4[\ell(0)]^{-1}\eta^{-1}V^2\varepsilon^{-1}$ iterations guarantees that $w = w_{T-1}$ satisfies

$$F_\ell(w) \leq F_\ell(v) + \ell(0) \cdot \varepsilon/3 + CVn^{-1/2},$$

for some absolute constant $C > 0$ depending only on L, B_X, and $\log(1/\delta)$. By Markov's inequality, this implies

$$\mathbb{P}(yw^\top x < 0) \leq \frac{1}{\ell(0)} F_\ell(w) \leq \frac{1}{\ell(0)} \left(F_\ell(v) + \frac{\ell(0)}{3}\varepsilon + CVn^{-1/2} \right). \tag{17}$$

Since $y\bar{v}^\top x \geq \gamma$ a.s., we have

$$F_\ell(v) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell(yV\bar{v}^\top x) \leq \ell(V\gamma).$$

If ℓ has polynomial tails, then by taking $V \geq \gamma^{-1}(6C_0[\ell(0)]^{-1}\varepsilon^{-1})^{1/p}$ we get $F_\ell(v) \leq C_0(\gamma V)^{-p} \leq \frac{\ell(0)\varepsilon}{6}$. Substituting this into (17), this implies

$$\mathbb{P}(yw^\top x < 0) \leq \frac{\varepsilon}{2} + \frac{CV}{\ell(0)n^{1/2}}. \tag{18}$$

Thus, provided $n = \Omega(\gamma^{-2}\varepsilon^{-2-\frac{2}{p}})$, if we run gradient descent for $T = \tilde{\Omega}(\gamma^{-2}\varepsilon^{-1-\frac{2}{p}})$ iterations, we have that $\mathrm{err}_{\mathcal{D}}^{0-1}(w) \leq \varepsilon$.

If ℓ has exponential tails, then by taking $V \geq \gamma^{-1}[C_1^{-1}\log(6C_0\ell(0)\varepsilon^{-1})]^{1/p}$ we get $F_\ell(v) \leq \frac{\ell(0)\varepsilon}{6}$, and so (18) holds in this case as well. This shows that for exponential tails, taking $n = \tilde{\Omega}(\gamma^{-2}\varepsilon^{-2})$ and $T = \tilde{\Omega}(\gamma^{-2}\varepsilon^{-1})$ suffices to achieve $\mathrm{err}_{\mathcal{D}}^{0-1}(w) \leq \varepsilon$. □

E. Remaining Proofs

In this section we provide the proof of Theorem 4.1. We first will prove the following bound on the empirical risk.

Lemma E.1. *Suppose that ℓ is convex and H-smooth. Assume $\|x\| \leq B_X$ a.s. Fix a step size $\eta \leq (2/5)H^{-1}B_X^{-2}$, and let $v \in \mathbb{R}^d$ be arbitrary. Then for any initialization w_0, and for any $\varepsilon > 0$, running gradient descent for $T = (4/3)\varepsilon^{-1}\eta^{-1}\|w_0 - v\|^2$ ensures that for all $t < T$, $\|w_t - v\| \leq \|w_0 - v\|$, and*

$$\widehat{F_\ell}(w_{T-1}) \leq \frac{1}{T}\sum_{t=0}^{T-1} \widehat{F_\ell}(w_t) \leq \widehat{F_\ell}(v) + \varepsilon.$$

To prove this, we first introduce the following upper bound for the norm of the gradient.

Lemma E.2 (Shamir (2020), Proof of Lemma 3)**.** *Suppose that ℓ is H-smooth. Then for any $\rho \in (0,1)$, provided $\eta \leq 2\rho H^{-1}B_X^{-2}$, $\widehat{F_\ell}(w_t)$ is decreasing in t. Moreover, if $T \in \mathbb{N}$ is arbitrary and $u \in \mathbb{R}^d$ is such that $\widehat{F_\ell}(u) \leq \widehat{F_\ell}(w_T)$, then for any $t < T$, we have the following gradient upper bound,*

$$\left\|\nabla\widehat{F_\ell}(w_t)\right\|^2 \leq \frac{1}{\eta(1-\rho)}\left(\widehat{F_\ell}(w_t) - \widehat{F_\ell}(u)\right). \tag{19}$$

With this gradient upper bound, we can prove Lemma E.1.

Proof. Let $\varepsilon > 0$ be fixed and let $T = (4/3)\varepsilon^{-1}\eta^{-1}\|w_0 - v\|^2$ be as in the statement of the lemma. We are done if $\widehat{F_\ell}(w_T) < \widehat{F_\ell}(v)$, so let us assume that $\widehat{F_\ell}(v) \leq \widehat{F_\ell}(w_T)$. We proceed by providing the appropriate lower bounds for

$$\|w_t - v\|^2 - \|w_{t+1} - v\|^2 = 2\eta\left\langle \widehat{F_\ell}(w_t), w_t - v\right\rangle - \eta^2\left\|\widehat{F_\ell}(w_t)\right\|^2.$$

For any $v \in \mathbb{R}^d$, by convexity of ℓ,

$$\begin{aligned}\left\langle \nabla \widehat{F_\ell}(w), w - v \right\rangle &= \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i w^\top x_i)(y_i w^\top x_i - y_i v^\top x_i) \\ &\geq \frac{1}{n}\sum_{i=1}^{n} [\ell(y_i w^\top x_i) - \ell(y_i v^\top x_i)] \\ &= \widehat{F_\ell}(w) - \widehat{F_\ell}(v), \end{aligned} \tag{20}$$

by convexity of ℓ. On the other hand, since $\widehat{F_\ell}(v) \leq \widehat{F_\ell}(w_T)$, by Lemma E.2, for any $t < T$, (19) holds, i.e.

$$\left\| \nabla \widehat{F_\ell}(w_t) \right\|^2 \leq \frac{1}{\eta(1-\rho)} \left(\widehat{F_\ell}(w_t) - \widehat{F_\ell}(v) \right). \tag{21}$$

Thus, for $\eta \leq (2/5) H^{-1} B_X^{-2}$, putting eqs. (20) and (21) together yields

$$\begin{aligned}\|w_t - v\|^2 - \|w_{t+1} - v\|^2 &= 2\eta \left\langle \nabla \widehat{F_\ell}(w_t), w_t - v \right\rangle - \eta^2 \left\| \nabla \widehat{F_\ell}(w_t) \right\|^2 \\ &\geq 2\eta(\widehat{F_\ell}(w_t) - \widehat{F_\ell}(v)) - \eta^2 \cdot \frac{1}{\eta(1-1/5)} \left(\widehat{F_\ell}(w_t) - \widehat{F_\ell}(v) \right) \\ &= \frac{3}{4}\eta \left(\widehat{F_\ell}(w_t) - \widehat{F_\ell}(v) \right). \end{aligned}$$

Summing and teloscoping over $t < T$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \widehat{F_\ell}(w_t) \leq \widehat{F_\ell}(v) + \frac{(4/3)\|w_0 - v\|^2}{\eta T} \leq \widehat{F_\ell}(v) + \varepsilon.$$

By Lemma E.2, $\widehat{F_\ell}(w_t)$ is decreasing in t, and therefore

$$\widehat{F_\ell}(w_{T-1}) = \min_{t<T} \widehat{F_\ell}(w_t) \leq T^{-1} \sum_{t<T} \widehat{F_\ell}(w_t),$$

completing the proof. □

Lemma E.1 shows that throughout the trajectory of gradient descent, $\|w_t\|$ stays bounded by the norm of the reference vector v. We can thus use Rademacher complexity bounds to prove Theorem 4.1.

Proof. By Lemma E.1, it suffices to show that the gap between the empirical and population surrogate risk is small. To do so, we use a Rademacher complexity argument. Denote by $\mathcal{G}$ the function class

$$\mathcal{G}_V := \{x \mapsto w^\top x : \|w\| \leq 3V\}.$$

Since ℓ is L-Lipschitz and $\ell(0) \leq 1$, it holds that $\ell(yw^\top x) \leq 1 + 3LV \leq 4LV$. We therefore use standard results in Rademacher complexity (e.g. Theorem 26.12 of Shalev-Shwartz & Ben-David, 2014) to get that with probability at least $1 - \delta$, for any $w \in \mathcal{G}_V$,

$$F_\ell(w) \leq \widehat{F_\ell}(w) + \frac{2B_X V L}{\sqrt{n}} + 4B_X V \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Since the output of gradient descent satisfies $\|w_{T-1} - v\| \leq \|w_0 - v\| \leq 2V$, we see that $w_{T-1} \in \mathcal{G}_V$. We can thus apply the Rademacher complexity bound to both $w_{T-1} \in \mathcal{G}_V$ and $v \in \mathcal{G}_V$, proving the theorem. □