Agnostic Learnability of Halfspaces via Logistic Loss

Ziwei Ji [1] **Kwangjun Ahn** [2] **Pranjal Awasthi** [3] **Satyen Kale** [3] **Stefani Karp** [3 4]

Abstract

We investigate approximation guarantees provided by logistic regression for the fundamental problem of agnostic learning of homogeneous halfspaces. Previously, for a certain broad class of "well-behaved" distributions on the examples, Diakonikolas et al. (2020d) proved an $\widetilde{\Omega}(\mathrm{OPT})$ lower bound, while Frei et al. (2021b) proved an $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound, where OPT denotes the best zero-one/misclassification risk of a homogeneous halfspace. In this paper, we close this gap by constructing a well-behaved distribution such that the global minimizer of the logistic risk over this distribution only achieves $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ misclassification risk, matching the upper bound in (Frei et al., 2021b). On the other hand, we also show that if we impose a radial-Lipschitzness condition in addition to well-behaved-ness on the distribution, logistic regression on a ball of bounded radius reaches $\widetilde{O}(\mathrm{OPT})$ misclassification risk. Our techniques also show for any well-behaved distribution, regardless of radial Lipschitzness, we can overcome the $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ lower bound for logistic loss simply at the cost of one additional convex optimization step involving the hinge loss and attain $\widetilde{O}(\mathrm{OPT})$ misclassification risk. 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.

1. Introduction

In this paper, we consider the fundamental problem of agnostically learning homogeneous halfspaces. Specifically, 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. A sample from P consists of an input feature vector $x \in \mathbb{R}^d$, and a binary label $y \in \{-1, +1\}$. Our goal is to compete with a homogeneous linear classifier $\bar{u}$ (i.e. one that predicts the label $\mathrm{sign}(\langle \bar{u}, x\rangle)$ for input x) that achieves the optimal zero-one risk of $\mathrm{OPT} > 0$ over P; formally, this means $\Pr_{(x,y)\sim P}\left(\mathrm{sign}(\langle \bar{u}, x\rangle) \neq y\right) = \mathrm{OPT}$. 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.

There have been many algorithmic and hardness results on this topic, see Section 1.1 for a discussion. 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: Ben-David et al. (2012) showed that the minimizer of the logistic risk may attain a zero-one risk as bad as $1 - \mathrm{OPT}$ on an adversarially-constructed distribution.

As a result, much attention has been devoted to certain "well-behaved" distributions, for which much better results can be obtained. However, even when the marginal distribution on the feature space, P_x, is assumed to be isotropic log-concave, in a recent work, Diakonikolas et al. (2020d) proved an $\widetilde{\Omega}\left(\mathrm{OPT}\right)$ lower bound on the zero-one risk for any convex surrogate, including the logistic loss. On the positive side, in another recent work, Frei et al. (2021b) proved that vanilla gradient descent on the logistic loss can attain a zero-one risk of $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$, as long as P_x satisfies some well-behavedness conditions. (See Sections 1.1 and 3 for precise details.)

The above results still leave a big gap between the upper and the lower bounds, raising the question of identifying the fundamental limits of logistic regression for this problem. In this work, we study this question and develop the following set of results.

Part of this work was done when Ziwei Ji and Kwangjun Ahn were interns at Google. [1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA [2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA [3]Google Research, New York City, NY, USA [4]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: Ziwei Ji <ziweiji2@illinois.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

A matching $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ lower bound. In Section 2, we 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 (Frei et al., 2021b), thereby closing the gap in recent works (Diakonikolas et al., 2020d; Frei et al., 2021b). Specifically, the marginal distribution Q_x is isotropic and bounded, and satisfies all the well-behaved-ness conditions from the aforementioned papers, but the global minimizer of the logistic risk on Q only attains $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ zero-one risk on Q.

An $\widetilde{O}(\mathrm{OPT})$ upper bound for radially Lipschitz densities. The lower bound mentioned above shows that one needs to make additional assumptions to prove better bounds. In Section 3, we show that by making a radial Lipschitzness assumption in addition to well-behaved-ness, it is indeed possible to achieve the near-optimal $\widetilde{O}(\mathrm{OPT})$ zero-one risk via logistic regression. In particular, our upper bound result holds if the projection of P_x onto any two-dimensional subspace has Lipschitz continuous densities. Moreover, our upper bound analysis is versatile: it can recover the $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ guarantee for general well-behaved distributions shown by Frei et al. (2021b), and it also works for the hinge loss, which motivates a simple and efficient two-phase algorithm, as we describe next.

An $\widetilde{O}(\mathrm{OPT})$ upper bound for general well-behaved distributions with a two-phase algorithm. Motivated by our analysis, in Section 4, we describe a simple two-phase algorithm that achieves $\widetilde{O}\left(\mathrm{OPT}\right)$ risk for general well-behaved distributions, without assuming radial Lipschitzness. Thus, we show that the cost of avoiding the radial Lipschitzness condition is simply an additional convex loss minimization. Our two-phase algorithm involves logistic regression followed by stochastic gradient descent with the hinge loss (i.e., the perceptron algorithm) with a restricted domain and a warm start. For general well-behaved distributions, the first phase can only achieve an $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ guarantee, however we show that the second phase can boost the upper bound to $\widetilde{O}(\mathrm{OPT})$.

Previously, for any given $\epsilon > 0$, Diakonikolas et al. (2020d) designed a nonconvex optimization algorithm that can achieve an $O(\mathrm{OPT} + \epsilon)$ risk using $\widetilde{O}(d/\epsilon^4)$ samples. Their algorithm requires guessing OPT within a constant multiplicative factor via a binary search and running a nonconvex SGD using each guess as an input. Similarly, prior algorithms achieving an $O(\mathrm{OPT} + \epsilon)$ risk involve solving multiple rounds of convex loss minimization (Awasthi et al., 2014; Daniely, 2015). In contrast, our two-phase algorithm is a simple logistic regression followed by a perceptron algorithm, and the output is guaranteed to have an $O\left(\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon\right)$ zero-one risk using only $\widetilde{O}(d/\epsilon^2)$ samples.

1.1. Related work

The problem of agnostic learning of halfspaces has a long and rich history (Kearns et al., 1994). 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 (Feldman et al., 2006; Guruswami & Raghavendra, 2009; Daniely, 2016). As a result, most algorithmic results have been obtained under assumptions on the marginal distribution P_x over the examples.

The work of Kalai et al. (2008) 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 a recent evidence that removing the exponential dependence on $1/\epsilon$, even for Gaussian marginals is computationally hard (Klivans & Kothari, 2014; Diakonikolas et al., 2020a; Goel et al., 2020).

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, Klivans et al. (2009) designed a polynomial-time algorithm that attains $\widetilde{O}(\mathrm{OPT}^{1/3}) + \epsilon$ zero-one risk for isotropic log-concave distributions. Awasthi et al. (2014) improved the upper bound to $O(\mathrm{OPT}) + \epsilon$, using a localization-based algorithm. Balcan & Zhang (2017) further extended the algorithm to more general s-concave distributions. The work of Daniely (2015) further provided a *PTAS* guarantee: an error of $(1 + \eta)\mathrm{OPT} + \epsilon$ for any desired constant $\eta > 0$ via an improper learner.

In a recent work, Diakonikolas et al. (2020d) studied the problem for distributions satisfying certain "well-behavedness" conditions which include isotropy and certain regularity conditions on the projection of P_x on any 2-dimensional subspace (see Assumption 3.2 for a subset of these conditions). This class of distributions include any isotropic log-concave distributions such as standard Gaussian. In addition to their nonconvex optimization method discussed above, for any convex, nonincreasing, and nonconstant loss function, they also showed 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.

In another recent work, Frei et al. (2021b) assumed P_x satisfies a "soft-margin" condition: for anti-concentrated marginals such as isotropic log-concave marginals, this assumes $\Pr\left(\left|\langle\bar{u}, x\rangle\right| \le \gamma\right) = O(\gamma)$ 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 (Diakonikolas et al., 2020d) do not match: if P_x is sub-exponential, then Diakonikolas et al. (2020d) only gave an $\widetilde{\Omega}(\mathrm{OPT})$ lower bound, while if P_x is s-heavy-tailed, then the upper bound in (Frei et al., 2021b) becomes worse.

Finally, some prior works on agnostic learning of halfspaces have considered various extensions of the problem such as active agnostic learning (Awasthi et al., 2014; Yan & Zhang, 2017), agnostic learning of sparse halfspaces with sample complexity scaling logarithmically in the ambient dimensionality (Shen & Zhang, 2021), and agnostic learning under weaker noise models such as the random classification noise (Blum et al., 1998; Dunagan & Vempala, 2008), Massart's noise model (Awasthi et al., 2015; 2016; Zhang et al., 2020; Diakonikolas et al., 2019; 2020b; 2021; Chen et al., 2020) and the Tsybakov noise model (Diakonikolas et al., 2020c; Zhang & Li, 2021). We do not consider these extensions in our work.

1.2. Notation

Let $\|\cdot\|$ denote the ℓ_2 (Euclidean) norm. Given $r > 0$, let $\mathcal{B}(r) := \{x \big| \|x\| \le r\}$ 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 $p_V(r, \theta)$.

Given a nonincreasing loss function $\ell : \mathbb{R} \to \mathbb{R}$, we consider the population risk

$$\mathcal{R}_\ell(w) := \mathbb{E}_{(x,y)\sim P}\left[\ell\left(y\langle w, x\rangle\right)\right],$$

and the corresponding empirical risk

$$\widehat{\mathcal{R}}_\ell(w) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i\langle w, x_i\rangle\right),$$

defined over n i.i.d. samples drawn from P. We will focus on the logistic loss $\ell_{\log}(z) := \ln(1 + e^{-z})$, and the hinge loss $\ell_h(z) := \max\{-z, 0\}$. Let $\mathcal{R}_{\log} := \mathcal{R}_{\ell_{\log}}$ for simplicity, and also define $\widehat{\mathcal{R}}_{\log}$, $\mathcal{R}_h$ and $\widehat{\mathcal{R}}_h$ similarly. Let $\mathcal{R}_{0-1}(w) := \Pr_{(x,y)\sim P}\left(y \ne \mathrm{sign}\left(\langle w, x\rangle\right)\right)$ denote the population zero-one risk.

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

In this section, we construct a distribution Q over $\mathbb{R}^2 \times \{-1, +1\}$ which satisfies standard regularity conditions in (Diakonikolas et al., 2020d; Frei et al., 2021a), 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 (Diakonikolas et al., 2020d); 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 $\mathrm{OPT} \le 1/16$, the construction is valid.

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 $\mathcal{B}(1) := \{x \big| \|x\| \le 1\}$ with density $q_4 := \frac{1 - \mathrm{OPT} - 2\sqrt{\mathrm{OPT}} - q_3}{\pi}$, and the label is given by $\mathrm{sign}(x_1)$.

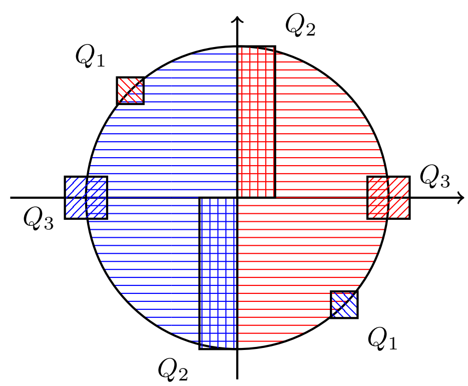


Figure 1. An illustration of Q when $\mathrm{OPT} = 1/16$. Red areas denote the $+1$ label, while blue areas denote the -1 label. The parts Q_1, Q_2 and Q_3 are marked in the figure, while Q_4 is supported on the unit circle and marked by horizontal lines.

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 the noisy part Q_1.

Here is our lower bound result.

Theorem 2.1. *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* $\mathbb{E}_{x\sim Q_x}[x_1x_2] = 0$*, 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*

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

Note that we can further normalize Q_x to unit variance and make it isotropic. Then it is easy to check that Q_x satisfies the "well-behaved-ness" conditions in (Diakonikolas et al., 2020d), and the "soft-margin" and "sub-exponential" conditions in (Frei et al., 2021b). In particular, our lower bound matches the upper bound in (Frei et al., 2021b).

2.1. Proof of Theorem 2.1

Here is a proof sketch of Theorem 2.1; the full proof is given in Appendix B.

First, basic calculation shows that Q_x is isotropic up to a constant multiplicative factor. Specifically, Q_1, Q_2 and Q_4 are constructed to make the risk lower bound proof work, while Q_3 is included to make Q isotropic. It turns out that Q_3 does not change the risk lower bound proof too much: the reason is that we will prove Theorem 2.1 by contradiction, and show that if its conclusion does not hold, then $\nabla\mathcal{R}_{\log}(w^*) \ne 0$. We show this mainly using Q_1, Q_2 and Q_4, but Q_3 does not significantly change the argument either, since it is highly aligned with the ground-truth solution $\bar{u} := (1, 0)$. As a result, if we assume the conclusion of Theorem 2.1 does not hold, which means Q_3 is also aligned with w^*, then $\ell'\left(y\langle w^*, x\rangle\right)$ will be close to 0 on Q_3, and we can still obtain a nonzero $\nabla\mathcal{R}_{\log}(w^*)$ and derive a contradiction.

Next we consider 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 a half of the distribution Q.

Note that the minimizer of the logistic risk on Q_4 by itself is infinitely far in the direction of $\bar{u}$. However, this will incur a large risk on Q_1. By balancing these two parts, we can 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 $O\left(\sqrt{\mathrm{OPT}}\right)$.

Lemma 2.2. *Suppose* $\mathrm{OPT} \le 1/100$*, let* $\bar{w} := (\bar{r}, 0)$ *where* $\bar{r} = \frac{3}{\sqrt{\mathrm{OPT}}}$*, then* $\mathcal{R}_{\log}(\bar{w}) \le 5\sqrt{\mathrm{OPT}}$*.*

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^*. We will assume $\theta^* \in \left[-\frac{\sqrt{\mathrm{OPT}}}{30}, \frac{\sqrt{\mathrm{OPT}}}{30}\right]$, and derive a contradiction.

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 2.2, a contradiction.

Lemma 2.3. *Suppose* $\mathrm{OPT} \le 1/100$ *and* $\theta^* \in \left[0, \frac{\sqrt{\mathrm{OPT}}}{30}\right]$*, then* $r^* \le \frac{10}{\sqrt{\mathrm{OPT}}}$*.*

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

To prove Lemma 2.4, let us consider an arbitrary $w = (r, \theta)$ under the conditions of Lemma 2.4. For simplicity, let us first look at the case $\theta = 0$. In this case, note that $y\langle w, x\rangle \le \sqrt{\mathrm{OPT}} \cdot \frac{10}{\sqrt{\mathrm{OPT}}} = 10$ on Q_2, therefore $\ell'\left(y\langle w, x\rangle\right)$ is bounded away from 0 on Q_2. We can then show that Q_2 induces a component of length $\frac{C_1}{\sqrt{\mathrm{OPT}}}$ in the gradient $\nabla\mathcal{R}_{\log}(w)$ along the direction of $-e_2 = (0, -1)$, where C_1 is some universal constant. Moreover, Q_1 also induces a component in the gradient along $-e_2$, while Q_3 and Q_4 induce a zero component along e_2. As a result, $\left\langle\nabla\mathcal{R}_{\log}(w), e_2\right\rangle < 0$, and thus $\nabla\mathcal{R}_{\log}(w)$ is nonzero. Now if $0 \le \theta \le C_2\sqrt{\mathrm{OPT}}$ for some small enough constant C_2 ($1/30$ in our case), we can show that Q_3 and Q_4 cannot cancel the effect of Q_2, and it still holds that $\left\langle\nabla\mathcal{R}_{\log}(w), e_2\right\rangle < 0$.

3. An $\widetilde{O}(\mathrm{OPT})$ upper bound for logistic loss with radial Lipschitzness

The $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ lower bound construction in Section 2 shows that further assumptions on the distribution are necessary in order to improve the upper bound on the zero-one risk of the logistic regression solution. In particular, we note that the distribution Q constructed in Section 2 has a *discontinuous* density. In this section, we show that if we simply add a very mild Lipschitz continuity condition on the density, then we can achieve $\widetilde{O}(\mathrm{OPT})$ zero-one risk using logistic regression.

First, we formally provide the standard assumptions from prior work. Because of the lower bound for s-heavy-tailed distributions from (Diakonikolas et al., 2020d) (cf. Section 1.1), to get an $\widetilde{O}(\mathrm{OPT})$ zero-one risk, we need to assume P_x has a light tail. Following (Frei et al., 2021b), we will either consider a bounded distribution, or assume P_x

is sub-exponential as defined below (cf. (Vershynin, 2018, 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 $t > 0$,

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

We also need the next assumption, which is part of the "well-behaved-ness" conditions from (Diakonikolas et al., 2020d).

Assumption 3.2. 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 $p_V(r, \theta) \ge 1/U$ for all $r \le R$, and $p_V(r, \theta) \le \sigma(r)$ for all $r \ge 0$, and $\int_0^\infty \sigma(r)\,\mathrm{d}r \le U$, and $\int_0^\infty r\sigma(r)\,\mathrm{d}r \le U$.

While Assumption 3.2 may look a bit technically involved, it basically consists of some mild concentration and anti-concentration conditions. In particular, for a broad class of distributions including isotropic log-concave distributions, the sub-exponential condition and Assumption 3.2 hold with α_1, α_2, U, R all being universal constants.

Finally, as discussed earlier, the previous conditions are also satisfied by Q from Section 2, and thus to get the improved $\widetilde{O}(\mathrm{OPT})$ risk bound, we need the following radial Lipschitz continuity assumption.

Assumption 3.3. There exists a measurable function $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ such that for any two-dimensional subspace V,

$$\left|p_V(r, \theta) - p_V(r, \theta')\right| \le \kappa(r)|\theta - \theta'|.$$

We will see Assumption 3.3 is crucial for the upper bound analysis in Lemma 3.13. For some concrete examples, note that if P_x is radially symmetric (e.g., standard Gaussian), then its projection onto any two-dimensional subspace V is also radially symmetric, therefore we can let $\kappa(r) = 0$. On the other hand, if p_V is λ-Lipschitz continuous on $\mathbb{R}^2$ under ℓ_2 (e.g., general Gaussian), then it implies $|p_V(r,\theta) - p_V(r,\theta')| \le \lambda r|\theta - \theta'|$, therefore we can let $\kappa(r) = \lambda r$.

Now we can state our main results. In the following, we denote the unit linear classifier with the optimal zero-one risk by $\bar{u}$, with $\mathcal{R}_{0-1}(\bar{u}) = \mathrm{OPT} \in (0, 1/e)$. Our first result shows that, with Assumption 3.3, minimizing the logistic risk yields a solution with $\widetilde{O}(\mathrm{OPT})$ zero-one risk.

Theorem 3.4. *Under Assumptions 3.2 and 3.3, let w^* denote the global minimizer of $\mathcal{R}_{\log}$.*

1. *If $\|x\| \le B$ almost surely, then*
$$\mathcal{R}_{0-1}(w^*) = O\left((1 + C_\kappa)\mathrm{OPT}\right),$$
where $C_\kappa := \int_0^B \kappa(r)\,\mathrm{d}r$.

2. *If P_x is (α_1, α_2)-sub-exponential, then*
$$\mathcal{R}_{0-1}(w^*) = O\left((1 + C_\kappa)\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})\right),$$
where $C_\kappa := \int_0^{3\alpha_2\ln(1/\mathrm{OPT})} \kappa(r)\,\mathrm{d}r$.

Remark 3.5. Given Theorem 3.4, we only need to estimate C_κ to get a concrete bound. First, for radially symmetric distributions, since $\kappa(r) = 0$, we have $C_\kappa = 0$. On the other hand, if p_V is λ-Lipschitz continuous on $\mathbb{R}^2$, then we can let $\kappa(r) = \lambda r$, and then by definition, we can show $C_\kappa \le \lambda B^2/2$ in the bounded case, and $C_\kappa \le 9\lambda\alpha_2^2\ln(1/\mathrm{OPT})^2/2$ in the sub-exponential case.

Theorem 3.4 shows that with radial Lipschitzness, the global minimizer can attain $\widetilde{O}(\mathrm{OPT})$ zero-one risk; next we also give an algorithmic result. Given a target error $\epsilon \in (0, 1)$, we consider projected gradient descent on the empirical risk with a norm bound of $1/\sqrt{\epsilon}$: let $w_0 := 0$, and

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

Our next result shows that projected gradient descent can also give an $\widetilde{O}(\mathrm{OPT})$ risk. Note that for the two cases discussed below (bounded or sub-exponential), we use the corresponding C_κ defined in Theorem 3.4.

Theorem 3.6. *Suppose Assumptions 3.2 and 3.3 hold.*

1. *If $\|x\| \le B$ almost surely, then with $\eta = 4/B^2$, and $O\left(\frac{\ln(1/\delta)}{\epsilon^4}\right)$ samples and $O\left(\frac{1}{\epsilon^{5/2}}\right)$ iterations, with probability $1-\delta$, projected gradient descent outputs w_t satisfying*
$$\mathcal{R}_{0-1}(w_t) = O\left((1 + C_\kappa)(\mathrm{OPT} + \epsilon)\right).$$

2. *On the other hand, if P_x is (α_1, α_2)-sub-exponential, then with $\eta = \widetilde{\Theta}(1/d)$, using $\widetilde{O}\left(\frac{d\ln(1/\delta)^3}{\epsilon^4}\right)$ samples and $\widetilde{O}\left(\frac{d\ln(1/\delta)^2}{\epsilon^{5/2}}\right)$ iterations, with probability $1-\delta$, projected gradient descent outputs w_t with*
$$\mathcal{R}_{0-1}(w_t) = O\left((1 + C_\kappa)(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon)\right).$$

Next we give proof outlines of our results; the full proofs are given in Appendix C. For simplicity, here we focus only on the bounded case, while the sub-exponential case will be handled in Appendix C. Although the proofs of the two cases share some similarity, we want to emphasize that the sub-exponential case does not follow by simply truncating the distribution to a certain radius and thus reducing to the bounded case. The reason is that the truncation radius can be as large as $\sqrt{d}$, while for the bounded case in our results, B is considered a constant independent of d and hidden in the O notation; therefore this truncation argument will introduce a $\mathrm{poly}(d)$ dependency in the final bound. By contrast, our zero-one risk upper bounds for sub-exponential distributinos only depend on α_1, α_2, U and R, but do not depend on d.

3.1. Proof of Theorems 3.4 and 3.6

Theorems 3.4 and 3.6 rely on the following key lemma which provides a zero-one risk bound on near optimal solutions to the logistic regression problem. It basically says that a near optimal solution with reasonably large norm also attains a good zero-one risk.

Lemma 3.7. *Under Assumptions 3.2 and 3.3, suppose* $\hat{w}$ *satisfies* $\mathcal{R}_{\log}(\hat{w}) \le \mathcal{R}_{\log}(\|\hat{w}\|\bar{u}) + \epsilon_\ell$ *for some* $\epsilon_\ell \in [0,1)$. *If* $\|x\| \le B$ *almost surely, then*

$$\mathcal{R}_{0-1}(\hat{w}) = O\left(\max\left\{\mathrm{OPT}, \sqrt{\tfrac{\epsilon_\ell}{\|\hat{w}\|}}, \tfrac{C_\kappa}{\|\hat{w}\|^2}\right\}\right).$$

In this subsection, we will sketch the proofs of Theorems 3.4 and 3.6 using Lemma 3.7; the details are given in Appendix C.1. In the next subsection, we will prove Lemma 3.7.

We first prove Theorem 3.4. Note that by Lemma 3.7, it suffices to show that $\|w^*\| = \Omega\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$ (since $\epsilon_\ell = 0$ in this case), which is true due to the next result.

Lemma 3.8. *Under Assumption 3.2, if* $\|x\| \le B$ *almost surely and* $\mathrm{OPT} < \frac{R^4}{200U^3B}$, *then* $\|w^*\| = \Omega\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$.

Next we prove Theorem 3.6. Once again motivated by Lemma 3.7, we need to show that projected gradient descent can achieve a near optimal logistic risk and a large norm. Recall that given the target (zero-one) error $\epsilon \in (0,1)$, we run projected gradient descent on a Euclidean ball with radius $1/\sqrt{\epsilon}$ (cf. Equation (1)). Using standard optimization and generalization analyses, we can prove the following guarantee on $\mathcal{R}_{\log}(w_t)$.

Lemma 3.9. *Let the target optimization error* $\epsilon_\ell \in (0,1)$ *and the failure probability* $\delta \in (0,1/e)$ *be given. If* $\|x\| \le B$ *almost surely, then with* $\eta = 4/B^2$, *using* $O\left(\frac{(B+1)^2\ln(1/\delta)}{\epsilon\epsilon_\ell^2}\right)$ *samples and* $O\left(\frac{B^2}{\epsilon\epsilon_\ell}\right)$ *iterations, with probability* $1-\delta$, *projected gradient descent outputs* w_t *satisfying*

$$\mathcal{R}_{\log}(w_t) \le \mathcal{R}_{\log}\left(\|w_t\|\bar{u}\right) + \epsilon_\ell. \tag{2}$$

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

Lemma 3.10. *Under Assumption 3.2, suppose*

$$\epsilon < \min\left\{\frac{R^4}{36U^2}, \frac{R^4}{72^2U^4}\right\} \quad \text{and} \quad \epsilon_\ell \le \sqrt{\epsilon},$$

and that Equation (2) holds. If $\|x\| \le B$ *almost surely and* $\mathrm{OPT} < \frac{R^4}{500U^3B}$, *then* $\|w_t\| = \Omega\left(\min\left\{\frac{1}{\sqrt{\epsilon}}, \frac{1}{\sqrt{\mathrm{OPT}}}\right\}\right)$.

Now to prove Theorem 3.6, we simply need to combine Lemmas 3.7, 3.9 and 3.10 with $\epsilon_\ell = \epsilon^{3/2}$.

3.2. Proof of Lemma 3.7

Here we give a proof sketch of Lemma 3.7; the details are given in Appendix C.2. As mentioned before, here we focus on the bounded setting; in the appendix, we also prove a version of Lemma 3.7 for the sub-exponential setting (cf. Lemma C.4). One remark is that some of the lemmas in the proof are also true for the hinge loss, and this fact will be crucial in the later discussion regarding our two-phase algorithm (cf. Section 4).

Let $\bar{w} := \|\hat{w}\|\bar{u}$, and consider $\ell \in \{\ell_{\log}, \ell_h\}$. 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

$$\mathcal{R}_\ell(\hat{w}) - \mathcal{R}_\ell(\bar{w}) - \mathbb{E}\left[\ell\Big(\mathrm{sign}\left(\langle\bar{w},x\rangle\right)\langle\hat{w},x\rangle\Big) - \ell\Big(\mathrm{sign}\left(\langle\bar{w},x\rangle\right)\langle\bar{w},x\rangle\Big)\right], \tag{3}$$

the second term is given by

$$\mathbb{E}\left[\ell\Big(\mathrm{sign}\left(\langle\bar{w},x\rangle\right)\langle\hat{w},x\rangle\Big) - \ell\Big(\mathrm{sign}\left(\langle\hat{w},x\rangle\right)\langle\hat{w},x\rangle\Big)\right], \tag{4}$$

and the third term is given by

$$\mathbb{E}\left[\ell\Big(\mathrm{sign}\left(\langle\hat{w},x\rangle\right)\langle\hat{w},x\rangle\Big) - \ell\Big(\mathrm{sign}\left(\langle\bar{w},x\rangle\right)\langle\bar{w},x\rangle\Big)\right], \tag{5}$$

where the expectations are taken over P_x.

We first bound term (3), 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 hinge loss, we have the following equality:

$$\text{term (3)} = \mathbb{E}\left[\mathbb{1}_{y \ne \mathrm{sign}\left(\langle\bar{w},x\rangle\right)} \cdot y\langle\bar{w}-\hat{w},x\rangle\right].$$

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 3.11. *For* $\ell \in \{\ell_{\log}, \ell_h\}$, *if* $\|x\| \le B$ *almost surely,*

$$|\text{term (3)}| \le B\|\bar{w}-\hat{w}\| \cdot \mathrm{OPT}.$$

Next we bound term (4). Note that we only need to consider the case where $\langle\bar{w},x\rangle$ and $\langle\hat{w},x\rangle$ have different signs; in this case, we can use the property $\ell(-z) - \ell(z) = z$ again and show the next result.

Lemma 3.12. *Under Assumption 3.2, for* $\ell \in \{\ell_{\log}, \ell_h\}$,

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

Lastly, we consider term (5). Note that it is 0 for the hinge loss ℓ_h, because $\ell_h(z) = 0$ when $z \ge 0$. For the logistic loss, term (5) is also 0 if P_x is radially symmetric; in general, we will bound it using Assumption 3.3.

Lemma 3.13. *For $\ell = \ell_h$, term (5) is 0. For $\ell = \ell_{\log}$, under Assumption 3.3, if $\|x\| \le B$ almost surely, then*

$$|\text{term (5)}| \le 12C_\kappa \cdot \varphi(\hat{w}, \bar{w})/\|\hat{w}\|,$$

where $C_\kappa := \int_0^B \kappa(r)\,\mathrm{d}r$.

Now we are ready to prove Lemma 3.7. For simplicity, here we let φ denote $\varphi(\hat{w}, \bar{w})$. For bounded distributions, Lemmas 3.11 to 3.13 imply

$$\begin{aligned} C_1\|\hat{w}\|\varphi^2 &\le \epsilon_\ell + B\|\bar{w} - \hat{w}\| \cdot \mathrm{OPT} + C_2 C_\kappa \cdot \varphi/\|\hat{w}\| \\ &\le \epsilon_\ell + B\|\hat{w}\|\varphi \cdot \mathrm{OPT} + C_2 C_\kappa \cdot \varphi/\|\hat{w}\|, \end{aligned}$$

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

1. $C_1\|\hat{w}\|\varphi^2 \le 3\epsilon_\ell$, which implies $\varphi = O\left(\sqrt{\epsilon_\ell/\|\hat{w}\|}\right)$;
2. $C_1\|\hat{w}\|\varphi^2 \le 3B\|\hat{w}\|\varphi \cdot \mathrm{OPT}$, and it follows that $\varphi = O(\mathrm{OPT})$;
3. $C_1\|\hat{w}\|\varphi^2 \le 3C_2C_\kappa \cdot \varphi/\|\hat{w}\|$, and it follows that $\varphi = O\left(C_\kappa/\|\hat{w}\|^2\right)$.

Therefore we can show a bound on the angle between $\bar{w}$ and $\hat{w}$, which further implies a zero-one risk bound for $\hat{w}$, in light of (Diakonikolas et al., 2020d, Claim 3.4) which is stated below.

Lemma 3.14. *Under Assumption 3.2,*

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

3.3. Recovering the general $\sqrt{\mathrm{OPT}}$ bound

Frei et al. (2021b) showed an $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound under the "soft-margin" and "sub-exponential" conditions. Here we give an alternative proof of this result using our proof technique. The result in this section will later serve as a guarantee of the first phase of our two-phase algorithm (cf. Section 4) that achieves $\widetilde{O}(\mathrm{OPT})$ risk.

Recall that the only place we need Assumption 3.3 is in the proof of Lemma 3.13. However, even without Assumption 3.3, we can still prove the following general bound which only needs Assumption 3.2.

Lemma 3.15. *Under Assumption 3.2, for $\ell = \ell_{\log}$,*

$$|\text{term (5)}| \le \frac{12U}{\|\hat{w}\|}.$$

Now with Lemma 3.15, we can prove a weaker but more general version of Lemma 3.7 (cf. Theorem C.15). Further invoking Lemmas 3.9 and 3.10 (cf. Lemmas C.6 and C.7 for the corresponding sub-exponential results), and let $\epsilon_\ell = \sqrt{\epsilon}$, we can show the next result. We present the bound in terms of the angle instead of zero-one risk for later applications in Section 4.

Lemma 3.16. *Given the target error $\epsilon \in (0,1)$ and the failure probability $\delta \in (0, 1/e)$, consider projected gradient descent (1). If $\|x\| \le B$ 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 gradient descent outputs w_t with*

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

On the other hand, 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 gradient descent outputs w_t with

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

The proofs of the results above are given in Appendix C.3.

4. An $\widetilde{O}(\mathrm{OPT})$ upper bound with hinge loss

We now show how to avoid Assumption 3.3 and achieve an $\widetilde{O}(\mathrm{OPT})$ zero-one risk bound using an extra step of hinge loss minimization. The key observation here is that the only place where Assumption 3.3 is used is in Lemma 3.13 for bounding term (5) for logistic loss. However, as noted in Lemma 3.13, for hinge loss, term (5) is conveniently 0. So a version of Lemma 3.7 holds for hinge loss, without using Assumption 3.3, and dropping the third term of $\frac{C_\kappa}{\|\hat{w}\|^2}$ in the max. Thus, to get an $\widetilde{O}(\mathrm{OPT})$ upper bound, it is enough to minimize the hinge loss to find a solution $\hat{w}$ such that $\|\hat{w}\| = \Omega(1)$ and $\mathcal{R}_h(\hat{w}) \le \mathcal{R}_h(\|\hat{w}\|\bar{u}) + \epsilon_\ell$ for some $\epsilon_\ell = \widetilde{O}((\mathrm{OPT}+\epsilon)^2)$. However, there is still one remaining challenge: note that the global minimizer of $\mathcal{R}_h$ is given by 0, while if we add the explicit requirement $\|\hat{w}\| = \Omega(1)$, the problem becomes nonconvex.

Fortunately, we can bypass having to solve this nonconvex problem by leveraging the solution of the logistic regression problem, which is guaranteed to make an angle of at most $\widetilde{O}(\sqrt{\mathrm{OPT}+\epsilon})$ with $\bar{u}$, even without Assumption 3.3, by Lemma 3.16. This solution, represented by a unit vector v, gives us a "warm start" for hinge loss minimization. Specifically, suppose we optimize the hinge loss over the halfspace

$$\mathcal{D} := \left\{ w \in \mathbb{R}^d \,\middle|\, \langle w, v\rangle \ge 1 \right\}, \tag{6}$$

then any solution we find must have norm at least 1. Furthermore, using the fact that $\varphi(v, \bar{u}) \le \widetilde{O}(\sqrt{\mathrm{OPT}+\epsilon})$

and the positive homogeneity of the hinge loss, we can also conclude that the optimizer of the hinge loss satisfies $\mathcal{R}_h(\hat{w}) \le \mathcal{R}_h(\|\hat{w}\|\bar{u}) + \epsilon_\ell$, giving us the desired solution.

While the above analysis does yield a simple two-phase polynomial time algorithm for getting an $\widetilde{O}(\mathrm{OPT})$ zero-one risk bound, closer analysis reveals a sample complexity requirement of $\widetilde{O}(1/\epsilon^4)$. We can improve the sample complexity requirement to $\widetilde{O}(1/\epsilon^2)$ by doing a custom analysis of SGD on the hinge loss (i.e., perceptron, (Novikoff, 1963)) inspired by the above considerations. Thus we get the following two-phase algorithm[1]:

1. Run projected gradient descent under the settings of Lemma 3.16, 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 Equation (6) starting from $w_0 := v$: at step t, we sample $(x_t, y_t) \sim P$, and let

$$w_{t+1} := \Pi_{\mathcal{D}}\left[w_t - \eta\ell_h'\left(y_t\langle w_t, x_t\rangle\right)y_t x_t\right]. \tag{7}$$

   Here, we set the convention that $\ell_h'(0) = -1$.

Below, we present the results regarding the expected zero-one risk for simplicity; we note that the results can be turned into high-probability bounds using the repeated probability amplification technique.

Theorem 4.1. *Given the target error $\epsilon \in (0, 1/e)$, suppose Assumption 3.2 holds.*

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

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

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

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

4.1. Proof of Theorem 4.1

Here we give a proof sketch of Theorem 4.1, and again, we focus on bounded distributions for simplicity. The full proof is given in Appendix D.

[1] Note that the parameters η, T, etc. in this section are all chosen for the second phase.

Let $\bar{r} := 1/\langle v, \bar{u}\rangle$, and thus $\bar{r}\bar{u} \in \mathcal{D}$; we will treat $\bar{r}\bar{u}$ as a reference solution in the proof. At step t, we have

$$\begin{aligned}&\|w_{t+1} - \bar{r}\bar{u}\|^2 \\ \le\ &\|w_t - \bar{r}\bar{u}\|^2 - 2\eta\left\langle \ell_h'\left(y_t\langle w_t, x_t\rangle\right)y_t x_t, w_t - \bar{r}\bar{u}\right\rangle \\ &+ \eta^2\ell_h'\left(y_t\langle w_t, x_t\rangle\right)^2\|x_t\|^2.\end{aligned} \tag{8}$$

Define

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

Taking expectation of Equation (8) w.r.t. (x_t, y_t), and note that $\|x\| \le B$ almost surely and $(\ell_h')^2 = -\ell_h'$, we have

$$\begin{aligned}&\mathbb{E}\left[\|w_{t+1} - \bar{r}\bar{u}\|^2\right] - \|w_t - \bar{r}\bar{u}\|^2 \\ \le&\ -2\eta\left\langle \nabla\mathcal{R}_h(w_t), w_t - \bar{r}\bar{u}\right\rangle + \eta^2 B^2\mathcal{M}(w_t) \\ \le&\ -2\eta\left(\mathcal{R}_h(w_t) - \mathcal{R}_h(\bar{r}\bar{u})\right) + \eta^2 B^2\mathcal{M}(w_t).\end{aligned} \tag{9}$$

To continue, we note the following lemma, which follows from Lemmas 3.11 to 3.13, and the homogeneity of the hinge loss ℓ_h.

Lemma 4.2. *Suppose Assumption 3.2 holds. Consider an arbitrary $w \in \mathcal{D}$, and let φ denote $\varphi(w, \bar{u})$. If $\|x\| \le B$ almost surely, then*

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

and

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

The remaining steps of the proof proceed as follows. We will prove the following: for $\varphi_t := \varphi(w_t, \bar{u})$,

$$\mathbb{E}\left[\min_{0\le t\le T}\varphi_t\right] = O(\mathrm{OPT}+\epsilon). \tag{10}$$

First, note that we can assume

$$\frac{2R^3}{3U\pi^2}\varphi_t \ge B\cdot\mathrm{OPT} \tag{11}$$

for all t, since otherwise Equation (10) holds vacuously. It then follows from Equation (11) and Lemma 4.2 that for $C_1 = 2R^3/(3U\pi^2)$,

$$\mathcal{R}_h(w_t) - \mathcal{R}_h(\|w_t\|\bar{u}) \ge C_1\|w_t\|\varphi_t^2 \ge C_1\varphi_t^2,$$

where we also use the fact that $\|w\| \ge 1$ for all $w \in \mathcal{D}$.

Next, note that $\mathcal{M}(w_t) = O(\varphi_t)$, due to Equation (11) and Lemma 3.14. If $\varphi_t \le \epsilon$, then Equation (10) also holds, otherwise we can assume $\epsilon \le \varphi_t$, and let $\eta = C_2\epsilon$ for some small enough constant C_2, such that

$$\eta B^2 \mathcal{M}(w_t) \le C_1 \epsilon \varphi_t \le C_1 \varphi_t^2.$$

Now Equation (9) and Lemma 4.2 imply

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

Taking the total expectation and telescoping the above inequality for all t, we have

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

Recall that

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

due to the first phase of the algorithm. Since $\eta = C_2\epsilon$, we can further let $T = \Omega(1/\epsilon^2)$ and finish the proof.

5. Open problems

We conclude our paper with some open questions. First, as shown by Theorem 4.1, we can achieve an $O\left(\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon\right)$ zero-one risk using the two-phase algorithm. However, previous algorithms attain an $O(\mathrm{OPT} + \epsilon)$ bound (Awasthi et al., 2014; Diakonikolas et al., 2020d). Is it possible to develop an algorithm that relies on solving a (small) constant number of convex problems and achieves an $O(\mathrm{OPT} + \epsilon)$ risk?

Next, it would be also interesting to extend our results to more practical neural network settings. On one hand, Frei et al. (2021a) showed that stochastic gradient descent on a two-layer leaky ReLU network of any width achieves an $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ zero-one risk, where OPT still denotes the best zero-one risk of a linear classifier. On the other hand, Ji et al. (2021) showed that a wide two-layer ReLU network can achieve the optimal Bayes risk; however, their results require the width of the network to depend on a complexity measure that could be exponentially large in the worst case. Can a neural network with a reasonable width reach a zero-one risk of $O(\mathrm{OPT})$?

Acknowledgements

Ziwei Ji thanks Matus Telgarsky for helpful discussions, and the NSF for support under grant IIS-1750051.

References

Awasthi, P., Balcan, M. F., and Long, P. M. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 449–458, 2014.

Awasthi, P., Balcan, M.-F., Haghtalab, N., and Urner, R. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory*, pp. 167–190. PMLR, 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*, pp. 152–192. PMLR, 2016.

Balcan, M.-F. and Zhang, H. Sample and computationally efficient learning algorithms under s-concave distributions. *arXiv preprint arXiv:1703.07758*, 2017.

Ben-David, S., Loker, D., Srebro, N., and Sridharan, K. Minimizing the misclassification error rate using a surrogate convex loss. *arXiv preprint arXiv:1206.6442*, 2012.

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

Bubeck, S. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Chen, S., Koehler, F., Moitra, A., and Yau, M. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. *arXiv preprint arXiv:2006.04787*, 2020.

Daniely, A. A ptas for agnostically learning halfspaces. In *Conference on Learning Theory*, pp. 484–502. PMLR, 2015.

Daniely, A. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 105–117, 2016.

Diakonikolas, I., Gouleakis, T., and Tzamos, C. Distribution-independent pac learning of halfspaces with massart noise. *arXiv preprint arXiv:1906.10075*, 2019.

Diakonikolas, I., Kane, D. M., and Zarifis, N. Near-optimal sq lower bounds for agnostically learning halfspaces and relus under gaussian marginals. *arXiv preprint arXiv:2006.16200*, 2020a.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory*, pp. 1486–1513. PMLR, 2020b.

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

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Non-convex sgd learns halfspaces with adversarial label noise. *arXiv preprint arXiv:2006.06742*, 2020d.

Diakonikolas, I., Kane, D. M., Kontonis, V., Tzamos, C., and Zarifis, N. Threshold phenomena in learning halfspaces with massart noise. *arXiv preprint arXiv:2108.08767*, 2021.

Dunagan, J. and Vempala, S. A simple polynomial-time rescaling algorithm for solving linear programs. *Mathematical Programming*, 114(1):101–114, 2008.

Feldman, V., Gopalan, P., Khot, S., and Ponnuswami, A. K. New results for learning noisy parities and halfspaces. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 563–574. IEEE, 2006.

Frei, S., Cao, Y., and Gu, Q. Provable generalization of sgd-trained neural networks of any width in the presence of adversarial label noise. *arXiv preprint arXiv:2101.01152*, 2021a.

Frei, S., Cao, Y., and Gu, Q. Agnostic learning of halfspaces with gradient descent via soft margins. In *International Conference on Machine Learning*, pp. 3417–3426. PMLR, 2021b.

Goel, S., Gollakota, A., and Klivans, A. Statistical-query lower bounds via functional gradients. *arXiv preprint arXiv:2006.15812*, 2020.

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

Ji, Z., Li, J. D., and Telgarsky, M. Early-stopped neural networks are consistent. *arXiv preprint arXiv:2106.05932*, 2021.

Kalai, A. T., Klivans, A. R., Mansour, Y., and Servedio, R. A. Agnostically learning halfspaces. *SIAM Journal on Computing*, 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.

Klivans, A. and Kothari, P. 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.

Klivans, A. R., Long, P. M., and Servedio, R. A. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12), 2009.

Novikoff, A. B. On convergence proofs for perceptrons. Technical report, STANFORD RESEARCH INST MENLO PARK CA, 1963.

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

Shen, J. and Zhang, C. Attribute-efficient learning of halfspaces with malicious noise: Near-optimal label complexity and noise tolerance. In *Algorithmic Learning Theory*, pp. 1072–1113. PMLR, 2021.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Yan, S. and Zhang, C. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. *arXiv preprint arXiv:1702.05581*, 2017.

Zhang, C. and Li, Y. Improved algorithms for efficient active learning halfspaces with massart and tsybakov noise. *arXiv preprint arXiv:2102.05312*, 2021.

Zhang, C., Shen, J., and Awasthi, P. Efficient active learning of sparse halfspaces with arbitrary bounded noise. *arXiv preprint arXiv:2002.04840*, 2020.

A. Technical lemmas

Here are some technical results we will need in our analysis.

Lemma A.1. *Let $r, \rho > 0$ be given, then*

$$\frac{2}{\rho}(1-e^{-r\rho}) \le \int_0^{2\pi} \ell_{\log}\left(r\rho\big|\cos(\theta)\big|\right) r\,\mathrm{d}\theta \le \frac{8\sqrt{2}}{\rho}.$$

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 $\sin(\theta) \ge \frac{\sqrt{2}}{2}$ for $\theta \in \left(\frac{\pi}{4}, \frac{\pi}{2}\right)$, therefore

$$4\int_0^{\frac{\pi}{2}} \ell_{\log}\left(r\rho\cos(\theta)\right) r\,\mathrm{d}\theta \le 8\int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \ell_{\log}\left(r\rho\cos(\theta)\right) r\,\mathrm{d}\theta \le \frac{8\sqrt{2}}{\rho}\int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \ell_{\log}\left(r\rho\cos(\theta)\right) r\rho\sin(\theta)\,\mathrm{d}\theta.$$

Also because $\ell_{\log}(z) \le \exp(-z)$,

$$\begin{aligned}
\int_0^{2\pi} \ell_{\log}\left(r\rho\big|\cos(\theta)\big|\right) r\,\mathrm{d}\theta &\le \frac{8\sqrt{2}}{\rho}\int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \exp\left(-r\rho\cos(\theta)\right) r\rho\sin(\theta)\,\mathrm{d}\theta \\
&= \frac{8\sqrt{2}}{\rho}\left(1-\exp\left(-\frac{\sqrt{2}r\rho}{2}\right)\right) \\
&\le \frac{8\sqrt{2}}{\rho}.
\end{aligned}$$

On the lower bound, note that $\ell_{\log}(z) \ge \frac{1}{2}\exp(-z)$ for $z \ge 0$, therefore

$$\begin{aligned}
\int_0^{2\pi} \ell_{\log}\left(r\rho\big|\cos(\theta)\big|\right) r\,\mathrm{d}\theta = 4\int_0^{\frac{\pi}{2}} \ell_{\log}\left(r\rho\cos(\theta)\right) r\,\mathrm{d}\theta &\ge 2\int_0^{\frac{\pi}{2}} \exp\left(-r\rho\cos(\theta)\right) r\,\mathrm{d}\theta \\
&\ge \frac{2}{\rho}\int_0^{\frac{\pi}{2}} \exp\left(-r\rho\cos(\theta)\right) r\rho\sin(\theta)\,\mathrm{d}\theta \\
&= \frac{2}{\rho}\left(1-e^{-r\rho}\right).
\end{aligned}$$

□

Lemma A.2. *Given $w, w' \in \mathbb{R}^d$, suppose $\Pr_{(x,y)\sim P}\Big(y \ne \mathrm{sign}\left(\langle w, x\rangle\right)\Big) = \mathrm{OPT}$. If $\|x\| \le B$ almost surely, then*

$$\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\ne\mathrm{sign}\left(\langle w,x\rangle\right)}\big|\langle w', x\rangle\big|\right] \le B\|w'\|\cdot\mathrm{OPT}.$$

If P_x is (α_1,α_2)-sub-exponential, and $\mathrm{OPT} \le \frac{1}{e}$, then

$$\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\ne\mathrm{sign}\left(\langle w,x\rangle\right)}\big|\langle w', x\rangle\big|\right] \le (1+2\alpha_1)\alpha_2\|w'\|\cdot\mathrm{OPT}\cdot\ln\left(\frac{1}{\mathrm{OPT}}\right).$$

Proof. If $\|x\| \le B$ almost surely, then

$$\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\ne\mathrm{sign}\left(\langle w,x\rangle\right)}\big|\langle w', x\rangle\big|\right] \le B\|w'\|\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\ne\mathrm{sign}\left(\langle w,x\rangle\right)}\right] = B\|w'\|\cdot\mathrm{OPT}.$$

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

$$\Pr\left(\left|\left\langle \frac{w'}{\|w'\|}, x\right\rangle\right| \geq t\right) \leq \alpha_1 \exp\left(-\frac{t}{\alpha_2}\right), \quad \text{equivalently} \quad \Pr\left(|\nu_x| \geq t\right) \leq \alpha_1 \exp\left(-\frac{t}{\alpha_2\|w'\|}\right).$$

Let $\mu(t) := \Pr\left(|\nu_x| \geq t\right)$. Given any threshold $\tau > 0$, integration by parts gives

$$\mathbb{E}\left[\mathbb{1}_{|\nu_x|\geq\tau}|\nu_x|\right] = \int_\tau^\infty t\cdot\left(-\,\mathrm{d}\mu(t)\right) = \tau\mu(\tau) + \int_\tau^\infty \mu(t)\,\mathrm{d}t \leq \alpha_1\left(\alpha_2\|w'\| + \tau\right)\exp\left(-\frac{\tau}{\alpha_2\|w'\|}\right). \tag{12}$$

Now let $\tau := \alpha_2\|w'\|\ln\left(\frac{1}{\mathrm{OPT}}\right)$. Note that

$$\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq\mathrm{sign}\left(\langle w,x\rangle\right)}\left|\langle w', x\rangle\right|\right] = \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{|\nu_x|\leq\tau}\mathbb{1}_{y\neq\mathrm{sign}\left(\langle w,x\rangle\right)}|\nu_x|\right] + \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{|\nu_x|\geq\tau}\mathbb{1}_{y\neq\mathrm{sign}\left(\langle w,x\rangle\right)}|\nu_x|\right].$$

We bound the two parts separately. When $|\nu_x| \leq \tau$, we have

$$\mathbb{E}\left[\mathbb{1}_{|\nu_x|\leq\tau}\mathbb{1}_{y\neq\mathrm{sign}\left(\langle w,x\rangle\right)}|\nu_x|\right] \leq \tau\mathbb{E}\left[\mathbb{1}_{y\neq\mathrm{sign}\left(\langle w,x\rangle\right)}\right] = \tau\cdot\mathrm{OPT} = \alpha_2\|w'\|\cdot\mathrm{OPT}\cdot\ln\left(\frac{1}{\mathrm{OPT}}\right).$$

On the other hand, when $|\nu_x| \geq \tau$, Equation (12) gives

$$\begin{aligned}\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{|\nu_x|\geq\tau}\mathbb{1}_{y\neq\mathrm{sign}\left(\langle w,x\rangle\right)}|\nu_x|\right] &\leq \mathbb{E}\left[\mathbb{1}_{|\nu_x|\geq\tau}|\nu_x|\right] \\ &\leq \alpha_1\alpha_2\|w'\|\left(1+\ln\left(\frac{1}{\mathrm{OPT}}\right)\right)\mathrm{OPT} \\ &\leq 2\alpha_1\alpha_2\|w'\|\cdot\mathrm{OPT}\cdot\ln\left(\frac{1}{\mathrm{OPT}}\right),\end{aligned}$$

where we also use $\mathrm{OPT} \leq \frac{1}{e}$. To sum up,

$$\mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq\mathrm{sign}\left(\langle w,x\rangle\right)}\left|\langle w', x\rangle\right|\right] \leq (1+2\alpha_1)\alpha_2\|w'\|\cdot\mathrm{OPT}\cdot\ln\left(\frac{1}{\mathrm{OPT}}\right).$$

□

B. Omitted proofs from Section 2

In this section, we will prove Theorem 2.1. First, we bound the density and support of Q_x.

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

Proof. For q_3, we have

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

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

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

therefore $q_4 \geq \frac{1}{2\pi}$. The upper bound $q_4 \leq \frac{1}{\pi}$ 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 $1+\frac{1}{\pi}\le 2$. □

Next we verify that Q_x is isotropic up to a multiplicative factor. We first note the following fact; its proof is straightforward and omitted.

Lemma B.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 the following result.

Lemma B.3. *It holds that* $\mathbb{E}_{x\sim Q_x}[x]=0$*, and* $\mathbb{E}_{x\sim Q_x}[x_1x_2]=0$*, 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 $\mathbb{E}_{x\sim Q_x}[x_1x_2]=0$, note that the expectation of x_1x_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 B.2, we have

$$\mathbb{E}_{(x,y)\sim Q_1}\left[x_1x_2\right]=-\frac{\mathrm{OPT}}{2}.$$

Additionally,

$$\mathbb{E}_{(x,y)\sim Q_2}\left[x_1x_2\right]=2\int_0^{\sqrt{\mathrm{OPT}}}\int_0^1 x_1x_2\,\mathrm{d}x_2\,\mathrm{d}x_1=\frac{\mathrm{OPT}}{2}.$$

Therefore $\mathbb{E}_{x\sim Q_x}[x_1x_2]=0$.

Finally, note that the expectation of $x_1^2-x_2^2$ is 0 on Q_1 due to Lemma B.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 B.2, it follows that $\mathbb{E}_{x\sim Q_x}\left[x_1^2-x_2^2\right]=0$. □

Next, we give a proof of the risk lower bound of Theorem 2.1. For simplicity, in this section we will let $\mathcal{R}$ denote $\mathcal{R}_{\log}$. For $i=1,2,3,4$, we also let $\mathcal{R}_i(w):=\mathbb{E}_{(x,y)\sim Q_i}\left[\ell_{\log}\left(y\langle w,x\rangle\right)\right]$; therefore $\mathcal{R}(w):=\sum_{i=1}^4\mathcal{R}_i(w)$. We first prove Lemma 2.2, showing that there exists a solution $\bar{w}$ with $\|\bar{w}\|=\Theta\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$ and $\mathcal{R}(\bar{w})=O\left(\sqrt{\mathrm{OPT}}\right)$.

Proof of Lemma 2.2. 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 B.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,

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

   where we use $q_4\le\frac{1}{\pi}$ from Lemma B.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

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

□

Next we prove Lemma 2.3, the upper bound on $\|w^*\|$.

Proof of Lemma 2.3. 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

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

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 2.2. Therefore $r^* \le \frac{10}{\sqrt{\mathrm{OPT}}}$. □

Next we prove Lemma 2.4.

Proof of Lemma 2.4. Let $w = (r, \theta)$, where $0 \le r \le \frac{10}{\sqrt{\mathrm{OPT}}}$ and $0 \le \theta \le \frac{\sqrt{\mathrm{OPT}}}{30}$. We will consider the projection of $\nabla\mathcal{R}(w)$ onto the direction $e_2 := (0, 1)$, and show that this projection cannot be zero.

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

$$2\int_0^{\sqrt{\mathrm{OPT}}}\int_0^1 \ell'_{\log}(x_1 w_1 + x_2 w_2)x_2 \,\mathrm{d}x_2\,\mathrm{d}x_1. \tag{13}$$

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

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

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

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

We can then upper bound Equation (13) as follows:

$$\begin{aligned}\text{Equation (13)} &\le 2\int_0^{\sqrt{\mathrm{OPT}}}\int_0^1 \ell'_{\log}\left(rx_1+\frac{1}{3}\right)x_2\,\mathrm{d}x_2\,\mathrm{d}x_1\\ &= \int_0^{\sqrt{\mathrm{OPT}}}\ell'_{\log}\left(rx_1+\frac{1}{3}\right)\mathrm{d}x_1\\ &= \frac{1}{r}\left(\ell_{\log}\left(\frac{1}{3}+r\sqrt{\mathrm{OPT}}\right)-\ell_{\log}\left(\frac{1}{3}\right)\right).\end{aligned}$$

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

$$\text{Equation (13)} \le \frac{1}{r}\ell'_{\log}\left(\frac{1}{3}+r\sqrt{\mathrm{OPT}}\right)r\sqrt{\mathrm{OPT}}\le \ell'_{\log}(3)\sqrt{\mathrm{OPT}}\le -\frac{\sqrt{\mathrm{OPT}}}{30}.$$

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

$$\text{Equation (13)} \le \frac{1}{r}\left(\ell_{\log}\left(\frac{7}{3}\right)-\ell_{\log}\left(\frac{1}{3}\right)\right)\le\frac{\sqrt{\mathrm{OPT}}}{10}\left(\ell_{\log}\left(\frac{7}{3}\right)-\ell_{\log}\left(\frac{1}{3}\right)\right)\le-\frac{\sqrt{\mathrm{OPT}}}{30}.$$

Therefore, it always holds that Equation (13) $\le -\frac{\sqrt{\mathrm{OPT}}}{30}$.

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 B.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

$$2\int_0^1\int_0^{2\theta}r'\cos(\theta')q_4r'\,\mathrm{d}\theta'\,\mathrm{d}r'=2q_4\cdot\frac{1}{3}\cdot\sin(2\theta)\le\frac{4\theta}{3\pi}\le\frac{4}{3\pi}\frac{\sqrt{\mathrm{OPT}}}{30}<\frac{\sqrt{\mathrm{OPT}}}{60},$$

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

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

Now we are ready to prove the risk lower bound of Theorem 2.1.

Proof of Theorem 2.1 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 $|\theta^*|\le\pi$. 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 2.3 then implies $r^* \leq \frac{10}{\sqrt{\mathrm{OPT}}}$, and then Lemma 2.4 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 B.1, we have

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

□

C. Omitted proofs from Section 3

In this section, we provide omitted proofs from Section 3. First, we prove some general results that will be used later.

Lemma C.1. *Under Assumption 3.2, for any $w \in \mathbb{R}^d$,*

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

Proof. 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

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

Invoking Assumption 3.2, we have

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

Lemma A.1 then implies

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

Then it follows from Assumption 3.2 that

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

□

Next, we note that following the direction of the ground-truth solution $\bar{u}$ can achieve $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ logistic risk.

Lemma C.2. *Given $\rho > 0$, under Assumption 3.2, if $\|x\| \leq B$ almost surely, then*

$$\mathcal{R}_{\log}(\rho\bar{u}) \leq \frac{12U}{\rho} + \rho B \cdot \mathrm{OPT}, \quad \text{with} \quad \inf_{\rho>0} \mathcal{R}_{\log}(\rho\bar{u}) \leq \sqrt{50UB \cdot \mathrm{OPT}},$$

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

$$\mathcal{R}_{\log}(\rho\bar{u}) \leq \frac{12U}{\rho} + (1+2\alpha_1)\alpha_2\rho \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right),$$

with

$$\inf_{\rho>0} \mathcal{R}_{\log}(\rho\bar{u}) \leq \sqrt{50(1+2\alpha_1)\alpha_2 U \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right)}.$$

Proof. Note that

$$\begin{aligned}\mathcal{R}_{\log}(\rho\bar{u}) &= \mathbb{E}_{(x,y)\sim P}\left[\ell_{\log}\left(y\langle\rho\bar{u},x\rangle\right)\right] \\ &= \mathbb{E}_{x\sim P_x}\left[\ell_{\log}\left(\left|\langle\rho\bar{u},x\rangle\right|\right)\right] + \mathbb{E}_{(x,y)\sim P}\left[\ell_{\log}\left(y\langle\rho\bar{u},x\rangle\right) - \ell_{\log}\left(\left|\langle\rho\bar{u},x\rangle\right|\right)\right].\end{aligned}$$

Since $\ell_{\log}(-z) - \ell_{\log}(z) = z$, and also invoking Lemma C.1, we have

$$\begin{aligned}\mathcal{R}_{\log}(\rho\bar{u}) &= \mathbb{E}_{x\sim P_x}\left[\ell_{\log}\left(\left|\langle\rho\bar{u},x\rangle\right|\right)\right] + \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq\mathrm{sign}\left(\langle\bar{u},x\rangle\right)}\cdot(-y)\langle\rho\bar{u},x\rangle\right] \\ &\le \frac{12U}{\rho} + \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y\neq\mathrm{sign}\left(\langle\bar{u},x\rangle\right)}\cdot(-y)\langle\rho\bar{u},x\rangle\right].\end{aligned}$$

If $\|x\| \le B$ almost surely, then Lemma A.2 further implies

$$\mathcal{R}_{\log}(\rho\bar{u}) \le \frac{12U}{\rho} + \rho B\cdot\mathrm{OPT},$$

and thus

$$\inf_{\rho>0}\mathcal{R}_{\log}(\rho\bar{u}) \le 2\sqrt{12UB\cdot\mathrm{OPT}} \le \sqrt{50UB\cdot\mathrm{OPT}}.$$

If P_x is (α_1,α_2)-sub-exponential, then Lemma A.2 further implies

$$\mathcal{R}_{\log}(\rho\bar{u}) \le \frac{12U}{\rho} + (1+2\alpha_1)\alpha_2\rho\cdot\mathrm{OPT}\cdot\ln\left(\frac{1}{\mathrm{OPT}}\right),$$

and therefore

$$\inf_{\rho>0}\mathcal{R}_{\log}(\rho\bar{u}) \le 2\sqrt{12(1+2\alpha_1)\alpha_2 U\cdot\mathrm{OPT}\cdot\ln\left(\frac{1}{\mathrm{OPT}}\right)} \le \sqrt{50(1+2\alpha_1)\alpha_2 U\cdot\mathrm{OPT}\cdot\ln\left(\frac{1}{\mathrm{OPT}}\right)}.$$

□

Next we prove a risk lower bound, that will later be used to prove lower bounds on $\|w^*\|$ and $\|w_t\|$.

Lemma C.3. *Under Assumption 3.2, given $w\in\mathbb{R}^d$, if $R\|w\|\le 2$, then*

$$\mathcal{R}_{\log}(w) \ge \frac{R^2}{2U},$$

while if $R\|w\|\ge 2$, then

$$\mathcal{R}_{\log}(w) \ge \frac{R}{U\|w\|}.$$

Proof. First, since $\ell_{\log}(z) \ge \ell_{\log}\left(|z|\right)$,

$$\mathcal{R}_{\log}(w) = \mathbb{E}_{(x,y)\sim P}\left[\ell_{\log}\left(y\langle w,x\rangle\right)\right] \ge \mathbb{E}_{x\sim P_x}\left[\ell_{\log}\left(\left|\langle w,x\rangle\right|\right)\right]. \tag{14}$$

Let v denote an arbitrary vector that is orthogonal to w, and let p denote the density of the projection of P_x onto the space spanned by w and v. Without loss of generality, we can assume w has polar angle 0. Then Equation (14) becomes

$$\mathcal{R}_{\log}(w) \ge \int_0^\infty\int_0^{2\pi}\ell_{\log}\left(r\|w\|\left|\cos(\theta)\right|\right)p(r,\theta)r\,\mathrm{d}\theta\,\mathrm{d}r.$$

Assumption 3.2 and Lemma A.1 then imply

$$\begin{aligned}\mathcal{R}_{\log}(w) &\geq \frac{1}{U}\int_0^R\int_0^{2\pi}\ell_{\log}\left(r\|w\|\left|\cos(\theta)\right|\right)r\,\mathrm{d}\theta\,\mathrm{d}r \\ &\geq \frac{1}{U}\frac{2}{\|w\|}\int_0^R\left(1-e^{-r\|w\|}\right)\mathrm{d}r \\ &= \frac{2}{U}\frac{1}{\|w\|^2}\left(e^{-R\|w\|}-1+R\|w\|\right).\end{aligned}$$

If $R\|w\|\leq 2$, then because $e^{-z}-1+z\geq\frac{z^2}{4}$ when $0\leq z\leq 2$, we have

$$\mathcal{R}_{\log}(w)\geq\frac{2}{U}\frac{1}{\|w\|^2}\frac{R^2\|w\|^2}{4}=\frac{R^2}{2U}.$$

Otherwise if $R\|w\|\geq 2$, then because $e^{-z}-1+z\geq\frac{z}{2}$ when $z\geq 2$, we have

$$\mathcal{R}_{\log}(w)\geq\frac{2}{U}\frac{1}{\|w\|^2}\frac{R\|w\|}{2}=\frac{R}{U\|w\|}.$$

□

C.1. Omitted proofs from Section 3.1

In this section, we prove Theorems 3.4 and 3.6. First, we state the following general version of Lemma 3.7, which also handles sub-exponential distributions; it will be proved in Appendix C.2.

Lemma C.4 (Lemma 3.7, including the sub-exponential case). *Under Assumptions 3.2 and 3.3, suppose $\hat{w}$ satisfies $\mathcal{R}_{\log}(\hat{w})\leq\mathcal{R}_{\log}(\|\hat{w}\|\bar{u})+\epsilon_\ell$ for some $\epsilon_\ell\in[0,1)$.*

1. *If $\|x\|\leq B$ almost surely, then*

$$\mathcal{R}_{0-1}(\hat{w})=O\left(\max\left\{\mathrm{OPT},\sqrt{\tfrac{\epsilon_\ell}{\|\hat{w}\|}},\tfrac{C_\kappa}{\|\hat{w}\|^2}\right\}\right).$$

2. *If P_x is (α_1,α_2)-sub-exponential and $\|\hat{w}\|=\Omega(1)$, then*

$$\mathcal{R}_{0-1}(\hat{w})= O\left(\max\left\{\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}),\sqrt{\tfrac{\epsilon_\ell}{\|\hat{w}\|}},\tfrac{C_\kappa}{\|\hat{w}\|^2}\right\}\right).$$

Next, we prove the following norm lower bound on $\|w^*\|$, which covers Lemma 3.8 and also the sub-exponential case.

Lemma C.5 (Lemma 3.8, including the sub-exponential case). *Under Assumption 3.2, if $\|x\|\leq B$ almost surely and $\mathrm{OPT}<\frac{R^4}{200U^3B}$, then $\|w^*\|=\Omega\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$; if P_x is (α_1,α_2)-sub-exponential and $\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})<\frac{R^4}{200(1+2\alpha_1)\alpha_2U^3}$, then $\|w^*\|=\Omega\left(\frac{1}{\sqrt{\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})}}\right)$.*

Proof. Suppose $\|x\|\leq B$ almost surely. Since $\mathrm{OPT}<\frac{R^4}{200U^3B}$, Lemma C.2 implies

$$\mathcal{R}_{\log}(w^*)\leq\inf_{\rho>0}\mathcal{R}_{\log}(\rho\bar{u})\leq\sqrt{50UB\cdot\mathrm{OPT}}<\sqrt{50UB\cdot\frac{R^4}{200U^3B}}=\frac{R^2}{2U}.$$

Therefore it follows from Lemma C.3 that $R\|w^*\|\geq 2$, and

$$\frac{R}{U\|w^*\|}\leq\mathcal{R}_{\log}(w^*)\leq\inf_{\rho>0}\mathcal{R}_{\log}(\rho\bar{u})\leq\sqrt{50UB\cdot\mathrm{OPT}},$$

which implies

$$\|w^*\| \geq \frac{R}{U\sqrt{50UB}} \cdot \frac{1}{\sqrt{\mathrm{OPT}}}.$$

Now suppose P_x is (α_1, α_2)-sub-exponential. Since $\mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right) < \frac{R^4}{200(1+2\alpha_1)\alpha_2 U^3}$, Lemma C.2 implies

$$\inf_{\rho>0} \mathcal{R}_{\log}(\rho\bar{u}) \leq \sqrt{50(1+2\alpha_1)\alpha_2 U \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right)} < \sqrt{50(1+2\alpha_1)\alpha_2 U \cdot \frac{R^4}{200(1+2\alpha_1)\alpha_2 U^3}} = \frac{R^2}{2U}.$$

Therefore it follows from Lemma C.3 that $R\|w^*\| \geq 2$, and

$$\frac{R}{U\|w^*\|} \leq \mathcal{R}_{\log}(w^*) \leq \inf_{\rho>0} \mathcal{R}_{\log}(\rho\bar{u}) \leq \sqrt{50(1+2\alpha_1)\alpha_2 U \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right)}$$

which implies

$$\|w^*\| \geq \frac{R}{U\sqrt{50(1+2\alpha_1)\alpha_2 U}} \frac{1}{\sqrt{\mathrm{OPT} \cdot \ln(1/\mathrm{OPT})}}.$$

□

Now we can prove Theorem 3.4.

Proof of Theorem 3.4. If $\|x\| \leq B$ almost surely, Lemma C.4 implies

$$\mathcal{R}_{0-1}(w^*) = O\left(\max\left\{\mathrm{OPT}, \frac{C_\kappa}{\|w^*\|^2}\right\}\right).$$

If $\mathrm{OPT} \geq \frac{R^4}{200U^3 B}$, then Theorem 3.4 holds vacuously; otherwise Lemma C.5 ensures $\|w^*\| = \Omega\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$, and thus

$$\mathcal{R}_{0-1}(w^*) = O\left(\max\left\{\mathrm{OPT}, C_\kappa \cdot \mathrm{OPT}\right\}\right) = O\left((1+C_\kappa)\mathrm{OPT}\right).$$

The proof of the sub-exponential case is similar. □

Next, we analyze project gradient descent. First we restate Lemmas 3.9 and 3.10, and also handle sub-exponential distributions.

Lemma C.6 (Lemma 3.9, including the sub-exponential case). *Let the target optimization error $\epsilon_\ell \in (0,1)$ and the failure probability $\delta \in (0, 1/e)$ be given. If $\|x\| \leq B$ almost surely, then with $\eta = 4/B^2$, using $O\left(\frac{(B+1)^2 \ln(1/\delta)}{\epsilon\epsilon_\ell^2}\right)$ samples and $O\left(\frac{B^2}{\epsilon\epsilon_\ell}\right)$ iterations, with probability $1-\delta$, projected gradient descent outputs w_t satisfying*

$$\mathcal{R}_{\log}(w_t) \leq \min_{0 \leq \rho \leq 1/\sqrt{\epsilon}} \mathcal{R}_{\log}(\rho\bar{u}) + \epsilon_\ell. \tag{15}$$

If P_x is (α_1, α_2)-sub-exponential, then with $\eta = \widetilde{\Theta}(1/d)$, using $\widetilde{O}\left(\frac{d\ln(1/\delta)^3}{\epsilon\epsilon_\ell^2}\right)$ samples and $\widetilde{O}\left(\frac{d\ln(1/\delta)^2}{\epsilon\epsilon_\ell}\right)$ iterations, with probability $1-\delta$, projected gradient descent outputs w_t satisfying Equation (15).

Lemma C.7 (Lemma 3.10, including the sub-exponential case). *Under Assumption 3.2, suppose*

$$\epsilon < \min\left\{\frac{R^4}{36U^2}, \frac{R^4}{72^2U^4}\right\} \quad \text{and} \quad \epsilon_\ell \leq \sqrt{\epsilon},$$

and that Equation (15) holds. If $\|x\| \leq B$ almost surely and $\mathrm{OPT} < \frac{R^4}{500U^3B}$, then $\|w_t\| = \Omega\left(\min\left\{\frac{1}{\sqrt{\epsilon}}, \frac{1}{\sqrt{\mathrm{OPT}}}\right\}\right)$.

On the other hand, if P_x is (α_1, α_2)-sub-exponential, and $\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) < \frac{R^4}{500U^3(1+2\alpha_1)\alpha_2}$, then it holds that $\|w_t\| = \Omega\left(\min\left\{\frac{1}{\sqrt{\epsilon}}, \frac{1}{\sqrt{\mathrm{OPT}\cdot\ln(1/\mathrm{OPT})}}\right\}\right)$.

Next we prove Lemmas C.6 and C.7. We first consider bounded distributions, and then handle sub-exponential distributions. For simplicity, in the rest of this subsection we will use $\mathcal{R}$ and $\widehat{\mathcal{R}}$ to denote $\mathcal{R}_{\log}$ and $\widehat{\mathcal{R}}_{\log}$, respectively.

Bounded distributions. First, here are some standard optimization and generalization results for projected gradient descent.

Lemma C.8. *If $\|x_i\| \le B$ for all $1 \le i \le n$, then $\widehat{\mathcal{R}}$ is $\frac{B^2}{4}$-smooth. Moreover, if $w_0 := 0$ and $\eta \le \frac{4}{B^2}$, then for all $t \ge 1$,*

$$\widehat{\mathcal{R}}(w_t) \le \min_{w \in \mathcal{B}(1/\sqrt{\epsilon})} \widehat{\mathcal{R}}(w) + \frac{1}{2\eta\epsilon t}.$$

Proof. Note that $\ell_{\log}$ is $\frac{1}{4}$-smooth. To show $\widehat{\mathcal{R}}$ is $\frac{B^2}{4}$-smooth, note that given any $w, w' \in \mathbb{R}^d$,

$$\begin{aligned}
\left\|\nabla\widehat{\mathcal{R}}(w) - \nabla\widehat{\mathcal{R}}(w')\right\| &= \left\|\frac{1}{n}\sum_{i=1}^n \Big(\ell'_{\log}\left(y_i\langle w, x_i\rangle\right) - \ell'_{\log}\left(y_i\langle w', x_i\rangle\right)\Big) y_i x_i\right\| \\
&\le \frac{1}{n}\sum_{i=1}^n \Big|\ell'_{\log}\left(y_i\langle w, x_i\rangle\right) - \ell'_{\log}\left(y_i\langle w', x_i\rangle\right)\Big| B \\
&\le \frac{B}{4n}\sum_{i=1}^n \big|y_i\langle w, x_i\rangle - y_i\langle w', x_i\rangle\big| \\
&\le \frac{B}{4n}\sum_{i=1}^n \|w - w'\| B = \frac{B^2}{4}\|w - w'\|.
\end{aligned}$$

The following analysis basically comes from the proof of (Bubeck, 2014, Theorem 6.3); we include it for completeness, and also handle the last iterate. Let $w^* := \arg\min_{w \in \mathcal{B}(1/\sqrt{\epsilon})} \widehat{\mathcal{R}}(w)$. Convexity gives

$$\widehat{\mathcal{R}}(w_t) - \widehat{\mathcal{R}}(w^*) \le \left\langle \nabla\widehat{\mathcal{R}}(w_t), w_t - w^* \right\rangle = \left\langle \nabla\widehat{\mathcal{R}}(w_t), w_t - w_{t+1} \right\rangle + \left\langle \nabla\widehat{\mathcal{R}}(w_t), w_{t+1} - w^* \right\rangle.$$

Smoothness implies

$$\begin{aligned}
\left\langle \nabla\widehat{\mathcal{R}}(w_t), w_t - w_{t+1} \right\rangle &\le \widehat{\mathcal{R}}(w_t) - \widehat{\mathcal{R}}(w_{t+1}) + \frac{B^2/4}{2}\|w_t - w_{t+1}\|^2 \\
&\le \widehat{\mathcal{R}}(w_t) - \widehat{\mathcal{R}}(w_{t+1}) + \frac{1}{2\eta}\|w_t - w_{t+1}\|^2.
\end{aligned}$$

On the other hand, the projection step ensures

$$\begin{aligned}
\left\langle \nabla\widehat{\mathcal{R}}(w_t), w_{t+1} - w^* \right\rangle &\le \frac{1}{\eta}\left\langle w_t - w_{t+1}, w_{t+1} - w^* \right\rangle \\
&= \frac{1}{2\eta}\Big(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2 - \|w_t - w_{t+1}\|^2\Big).
\end{aligned}$$

Therefore

$$\widehat{\mathcal{R}}(w_t) - \widehat{\mathcal{R}}(w^*) \le \widehat{\mathcal{R}}(w_t) - \widehat{\mathcal{R}}(w_{t+1}) + \frac{1}{2\eta}\Big(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2\Big),$$

which implies

$$\widehat{\mathcal{R}}(w_{t+1}) - \widehat{\mathcal{R}}(w^*) \le \frac{1}{2\eta}\Big(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2\Big). \tag{16}$$

Next we show that $\widehat{\mathcal{R}}(w_{t+1}) \le \widehat{\mathcal{R}}(w_t)$. Smoothness implies

$$\begin{aligned}
\widehat{\mathcal{R}}(w_{t+1}) - \widehat{\mathcal{R}}(w_t) &\le \left\langle \nabla\widehat{\mathcal{R}}(w_t), w_{t+1} - w_t \right\rangle + \frac{B^2/4}{2}\|w_{t+1} - w_t\|^2 \\
&\le -\frac{1}{\eta}\|w_t - w_{t+1}\|^2 + \frac{B^2/4}{2}\|w_{t+1} - w_t\|^2 \\
&\le -\frac{1}{\eta}\|w_t - w_{t+1}\|^2 + \frac{1}{2\eta}\|w_{t+1} - w_t\|^2 \\
&= -\frac{1}{2\eta}\|w_{t+1} - w_t\|^2,
\end{aligned}$$

where we also use the property of the projection step on the second line.

It now follow from Equation (16) and $\widehat{\mathcal{R}}(w_{t+1}) \le \widehat{\mathcal{R}}(w_t)$ that for $t \ge 1$,

$$\widehat{\mathcal{R}}(w_t) \le \widehat{\mathcal{R}}(w^*) + \frac{\|w_0 - w^*\|^2}{2\eta t} \le \widehat{\mathcal{R}}(w^*) + \frac{1}{2\eta\epsilon t}.$$

□

Lemma C.9. *If $\|x\| \le B$ almost surely, then with probability $1-\delta$, for all $w \in \mathcal{B}\left(\frac{1}{\sqrt{\epsilon}}\right)$,*

$$\left|\mathcal{R}(w) - \widehat{\mathcal{R}}(w)\right| \le \frac{2B}{\sqrt{\epsilon n}} + 3\left(\frac{B}{\sqrt{\epsilon}} + 1\right)\sqrt{\frac{\ln(4/\delta)}{2n}}.$$

Proof. Note that $\ell_{\log}(z) \le |z| + 1$, therefore

$$\ell_{\log}\left(y\langle w, x\rangle\right) \le \|w\|\|x\| + 1 \le \frac{B}{\sqrt{\epsilon}} + 1.$$

Since $\ell_{\log}$ is 1-Lipschitz continuous, (Shalev-Shwartz & Ben-David, 2014, Theorem 26.5, Lemma 26.9, Lemma 26.10) imply that with probability $1-\delta$, for all $w \in \mathcal{B}\left(\frac{1}{\sqrt{\epsilon}}\right)$,

$$\mathcal{R}(w) - \widehat{\mathcal{R}}(w) \le \frac{2B}{\sqrt{\epsilon n}} + 3\left(\frac{B}{\sqrt{\epsilon}} + 1\right)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Next we can just apply the same technique and get a uniform deviation bound on $\widehat{\mathcal{R}}(w) - \mathcal{R}(w)$. □

We can now prove Lemma C.6.

Proof of Lemma C.6 for bounded distributions. Lemma C.8 implies that

$$\widehat{\mathcal{R}}(w_t) - \min_{0\le\rho\le 1/\sqrt{\epsilon}} \mathcal{R}(\rho\bar{u}) \le \frac{1}{2\eta\epsilon t} = \frac{B^2}{8\epsilon t}.$$

Moreover, Lemma C.9 ensures with probability $1-\delta$, for all $w \in \mathcal{B}_2\left(\frac{1}{\sqrt{\epsilon}}\right)$,

$$\left|\widehat{\mathcal{R}}(w) - \mathcal{R}(w)\right| \le \frac{2B}{\sqrt{\epsilon n}} + 3\left(\frac{B}{\sqrt{\epsilon}} + 1\right)\sqrt{\frac{\ln(4/\delta)}{2n}} = O\left((B+1)\sqrt{\frac{\ln(1/\delta)}{\epsilon n}}\right).$$

Therefore, to ensure $\mathcal{R}(w_t) - \min_{0\le\rho\le 1/\sqrt{\epsilon}} \mathcal{R}(\rho\bar{u}) \le \epsilon_\ell$, we only need

$$O\left(\frac{B^2}{\epsilon\epsilon_\ell}\right) \text{ steps, } \quad \text{and} \quad O\left(\frac{(B+1)^2\ln(1/\delta)}{\epsilon\epsilon_\ell^2}\right) \text{ samples.}$$

□

Next we prove the norm lower bound on $\|w_t\|$.

Proof of Lemma C.7. First, we consider the case $\|x\| \le B$ almost surely. It follows from Lemma C.2 that

$$\mathcal{R}(\rho\bar{u}) \le \frac{12U}{\rho} + \rho B \cdot \mathrm{OPT}. \tag{17}$$

Let $\bar{\rho} := \sqrt{\frac{12U}{B \cdot \mathrm{OPT}}}$. We consider two cases below, $\bar{\rho} \le \frac{1}{\sqrt{\epsilon}}$ or $\bar{\rho} \ge \frac{1}{\sqrt{\epsilon}}$.

First, we assume $\bar{\rho} \le \frac{1}{\sqrt{\epsilon}}$. Then by the conditions of Lemma C.7 and Equation (17), we have

$$\begin{aligned}\mathcal{R}(w_t) \le \mathcal{R}(\bar{\rho}\bar{u}) + \epsilon_\ell &\le 2\sqrt{12UB \cdot \mathrm{OPT}} + \sqrt{\epsilon} \\ &< 2\sqrt{12UB \cdot \frac{R^4}{500U^3B}} + \sqrt{\frac{R^4}{36U^2}} \\ &< 2\frac{R^2}{6U} + \frac{R^2}{6U} = \frac{R^2}{2U}.\end{aligned}$$

It then follows from Lemma C.3 that $R\|w_t\| \ge 2$, and

$$\frac{R}{U\|w_t\|} \le \mathcal{R}(w_t) \le \mathcal{R}(\bar{\rho}\bar{u}) + \epsilon_\ell \le 2\sqrt{12UB \cdot \mathrm{OPT}} + \sqrt{\epsilon}.$$

since $\bar{\rho} \le \frac{1}{\sqrt{\epsilon}}$,

$$\sqrt{\epsilon} \le \frac{1}{\bar{\rho}} = \sqrt{\frac{B \cdot \mathrm{OPT}}{12U}}.$$

As a result, $\frac{R}{U\|w_t\|} = O\left(\sqrt{\mathrm{OPT}}\right)$, which implies $\|w_t\| = \Omega\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$.

Next, assume $\bar{\rho} \ge \frac{1}{\sqrt{\epsilon}}$, which implies that

$$\frac{B \cdot \mathrm{OPT}}{12U} \le \epsilon, \quad \text{and} \quad B \cdot \mathrm{OPT} \le 12U\epsilon.$$

Moreover, Equation (17) implies

$$\mathcal{R}\left(\frac{1}{\sqrt{\epsilon}}\bar{u}\right) \le 12U\sqrt{\epsilon} + \frac{1}{\sqrt{\epsilon}}B \cdot \mathrm{OPT} \le 12U\sqrt{\epsilon} + \frac{1}{\sqrt{\epsilon}}12U\epsilon = 24U\sqrt{\epsilon}.$$

Then because

$$\mathcal{R}(w_t) \le \mathcal{R}\left(\frac{1}{\sqrt{\epsilon}}\bar{u}\right) + \epsilon_\ell \le 24U\sqrt{\epsilon} + \sqrt{\epsilon} < 24U\sqrt{\frac{R^4}{72^2U^4}} + \sqrt{\frac{R^4}{36U^2}} = \frac{R^2}{2U},$$

it further follows from Lemma C.3 that $R\|w_t\| \ge 2$, and

$$\frac{R}{U\|w_t\|} \le \mathcal{R}\left(\frac{1}{\sqrt{\epsilon}}\bar{u}\right) + \epsilon_\ell \le 24U\sqrt{\epsilon} + \sqrt{\epsilon},$$

therefore $\|w_t\| = \Omega\left(\frac{1}{\sqrt{\epsilon}}\right)$.

Now assume P_x is (α_1, α_2)-sub-exponential. Lemma C.2 implies

$$\mathcal{R}(\rho\bar{u}) \le \frac{12U}{\rho} + (1 + 2\alpha_1)\alpha_2\rho \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right).$$

Let

$$\bar{\rho} := \sqrt{\frac{12U}{(1 + 2\alpha_1)\alpha_2 \cdot \mathrm{OPT} \cdot \ln(1/\mathrm{OPT})}},$$

and similarly consider the two cases $\bar{\rho} \le \frac{1}{\sqrt{\epsilon}}$ and $\bar{\rho} \ge \frac{1}{\sqrt{\epsilon}}$, we can finish the proof. □

Now we are ready to prove Theorem 3.6.

Proof of Theorem 3.6 for bounded distributions. First, note that if ϵ or OPT does not satisfy the conditions of Lemma C.7, then Theorem 3.6 holds vacuously. Under the conditions of Lemmas C.6 and C.7, let $\epsilon_\ell := \epsilon^{3/2}$, we have that projected gradient descent can find w_t satisfying

$$\mathcal{R}_{\log}(w_t) \le \min_{0\le\rho\le 1/\sqrt{\epsilon}} \mathcal{R}_{\log}(\rho\bar{u}) + \epsilon^{3/2} \le \mathcal{R}_{\log}\left(\|w_t\|\bar{u}\right) + \epsilon^{3/2},$$

and

$$\|w_t\| = \Omega\left(\min\left\{\frac{1}{\sqrt{\epsilon}}, \frac{1}{\sqrt{\mathrm{OPT}}}\right\}\right).$$

Now we just need to invoke Lemma C.4. If $\epsilon \le \mathrm{OPT}$, then $\|w_t\| = \Omega\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$, and Lemma C.4 implies

$$\begin{aligned}&\mathcal{R}_{0-1}(w_t)\\ &= O\left(\max\left\{\mathrm{OPT}, \sqrt{\epsilon^{3/2}\sqrt{\mathrm{OPT}}}, C_\kappa \cdot \mathrm{OPT}\right\}\right)\\ &= O\left((1+C_\kappa)\mathrm{OPT}\right).\end{aligned}$$

If $\epsilon \ge \mathrm{OPT}$, then $\|w_t\| = \Omega\left(\frac{1}{\sqrt{\epsilon}}\right)$, and similarly we can show

$$\begin{aligned}&\mathcal{R}_{0-1}(w_t)\\ &= O\left(\max\left\{\mathrm{OPT}, \sqrt{\epsilon^{3/2}\sqrt{\epsilon}}, C_\kappa \epsilon\right\}\right)\\ &= O\left((1+C_\kappa)(\mathrm{OPT}+\epsilon)\right).\end{aligned}$$

The sample and iteration complexity follow from Lemma C.6 and that $\epsilon_\ell = \epsilon^{3/2}$. □

Sub-exponential distributions. Next we handle (α_1, α_2)-sub-exponential distributions. We will prove Lemma C.6 for sub-exponential distributions; the rest of the proof is similar to the bounded case and thus omitted.

Let the target zero-one error ϵ, the target optimization error ϵ_ℓ, and failure probability δ be given. Given $r > 0$, we overload the notation a little bit and let

$$\delta(r) := d\alpha_1 \exp\left(-\frac{r}{\alpha_2\sqrt{d}}\right).$$

In particular, note that

$$\Pr_{x\sim P_x}\left(\|x\| \ge r\right) \le \sum_{j=1}^{d} \Pr\left(|x_j| \ge \frac{r}{\sqrt{d}}\right) \le d\alpha_1 \exp\left(-\frac{r}{\sqrt{d}\alpha_2}\right) = \delta(r).$$

Let $B > 1$ be large enough such that

$$\left(1-\delta(B)\right)^{100(B+1)^2\ln(4/\delta)/(\epsilon\epsilon_\ell^2)} \ge 1-\delta, \quad \text{and} \quad \alpha_1(\alpha_2+B)\exp\left(-\frac{B}{\alpha_2}\right) \le \epsilon_\ell\sqrt{\epsilon}. \tag{18}$$

We have the following bound on B.

Lemma C.10. *To satisfy Equation (18), it is enough to let*

$$B = \Omega\left(\sqrt{d}\ln\left(\frac{d}{\epsilon\epsilon_\ell\delta}\right)\right).$$

Proof. First, we let $B \geq \alpha_2\sqrt{d}\ln(2d\alpha_1)$ to ensure $\delta(B) \leq 1/2$. Since for $0 \leq z \leq 1/2$, we have $e^{-z} \geq 1 - z \geq e^{-2z}$, to satisfy the first condition of Equation (18), it is enough to ensure

$$e^{-\delta(B)\cdot 200(B+1)^2\ln(4/\delta)/(\epsilon\epsilon_\ell^2)} \geq e^{-\delta}, \quad \text{equivalently} \quad \delta(B) \leq \frac{\delta\epsilon\epsilon_\ell^2}{200(B+1)^2\ln(4/\delta)}.$$

Invoking the definition of $\delta(B)$, we only need

$$B \geq \alpha_2\sqrt{d}\ln\left(\frac{200(B+1)^2 d\alpha_1\ln(4/\delta)}{\delta\epsilon\epsilon_\ell^2}\right).$$

In other words, it is enough if $B = \Omega\left(\sqrt{d}\ln\left(\frac{d}{\epsilon\epsilon_\ell\delta}\right)\right)$.

Similarly, to satisfy the second condition of Equation (18), we only need

$$B \geq \alpha_2\ln\left(\frac{\alpha_1(\alpha_2+B)}{\epsilon_\ell\sqrt{\epsilon}}\right),$$

and it is enough if $B = \Omega\left(\sqrt{d}\ln\left(\frac{d}{\epsilon\epsilon_\ell\delta}\right)\right)$. □

Now we define a truncated logistic loss $\ell_{\log}^\circ$ as following:

$$\ell_{\log}^\circ(z) := \begin{cases} \ell_{\log}\left(-\frac{B}{\sqrt{\epsilon}}\right) & \text{if } z \leq -\frac{B}{\sqrt{\epsilon}}, \\ \ell_{\log}(z) & \text{if } z \geq -\frac{B}{\sqrt{\epsilon}}. \end{cases}$$

We also let $\mathcal{R}^\circ(w)$ and $\widehat{\mathcal{R}}^\circ(w)$ denote the population and empirical risk with the truncated logistic loss. We have the next result.

Lemma C.11. *Suppose $B > 1$ is chosen according to Equation (18). Using a constant step size $4/B^2$, and*

$$\frac{100(B+1)^2\ln(4/\delta)}{\epsilon\epsilon_\ell^2} \text{ samples}, \quad \text{and} \quad \frac{B^2}{4\epsilon\epsilon_\ell} \text{ steps},$$

with probability $1-2\delta$, projected gradient descent can ensure

$$\mathcal{R}^\circ(w_t) \leq \min_{0\leq\rho\leq 1/\sqrt{\epsilon}} \mathcal{R}(\rho\bar{u}) + \epsilon_\ell.$$

Proof. It follows from Equation (18) that with probability $1-\delta$, it holds that $\|x_i\| \leq B$ for all training examples. Therefore Lemma C.8 implies that

$$\widehat{\mathcal{R}}(w_t) \leq \min_{0\leq\rho\leq 1/\sqrt{\epsilon}} \widehat{\mathcal{R}}(\rho\bar{u}) + \frac{B^2}{8\epsilon t}.$$

Since $\|x_i\| \leq B$, and the domain is $\mathcal{B}(1/\sqrt{\epsilon})$, it follows that

$$\widehat{\mathcal{R}}^\circ(w_t) \leq \min_{0\leq\rho\leq 1/\sqrt{\epsilon}} \widehat{\mathcal{R}}^\circ(\rho\bar{u}) + \frac{B^2}{8\epsilon t}.$$

Letting $t = \frac{B^2}{4\epsilon\epsilon_\ell}$, we get

$$\widehat{\mathcal{R}}^\circ(w_t) \leq \min_{0\leq\rho\leq 1/\sqrt{\epsilon}} \widehat{\mathcal{R}}^\circ(\rho\bar{u}) + \frac{\epsilon_\ell}{2}. \tag{19}$$

Note that by the construction of the truncated logistic loss, it holds that

$$\ell^{\circ}_{\log}(z) \le \frac{B}{\sqrt{\epsilon}} + 1.$$

Then by invoking the standard Rademacher complexity results (Shalev-Shwartz & Ben-David, 2014, Theorem 26.5, Lemma 26.9, Lemma 26.10), and recall that we work under the event $\|x_i\| \le B$ for all training examples, we can show with probability $1-2\delta$ that for all $w \in \mathcal{B}(1/\sqrt{\epsilon})$,

$$\begin{aligned}\left|\mathcal{R}^{\circ}(w) - \widehat{\mathcal{R}}^{\circ}(w)\right| &\le \frac{2B}{\sqrt{\epsilon n}} + 3\left(\frac{B}{\sqrt{\epsilon}} + 1\right)\sqrt{\frac{\ln(4/\delta)}{2n}} \\ &\le \frac{2(B+1)}{\sqrt{\epsilon}}\sqrt{\frac{\ln(4/\delta)}{n}} + \frac{3(B+1)}{\sqrt{\epsilon}}\sqrt{\frac{\ln(4/\delta)}{2n}} \\ &\le 5(B+1)\sqrt{\frac{\ln(4/\delta)}{\epsilon n}}.\end{aligned}$$

Letting $n = \frac{100(B+1)^2\ln(4/\delta)}{\epsilon\epsilon_\ell^2}$, we have

$$\left|\mathcal{R}^{\circ}(w) - \widehat{\mathcal{R}}^{\circ}(w)\right| \le \frac{\epsilon_\ell}{2}. \tag{20}$$

It then follows from Equations (19) and (20) that with probability $1-2\delta$,

$$\mathcal{R}^{\circ}(w_t) \le \min_{0\le\rho\le 1/\sqrt{\epsilon}} \mathcal{R}^{\circ}(\rho\bar{u}) + \epsilon_\ell \le \min_{0\le\rho\le 1/\sqrt{\epsilon}} \mathcal{R}(\rho\bar{u}) + \epsilon_\ell,$$

where we use $\ell^{\circ}_{\log} \le \ell_{\log}$ in the last inequality. □

Finally, we show that $\mathcal{R}^{\circ}(w_t)$ is close to $\mathcal{R}(w_t)$.

Lemma C.12. *For all $w \in \mathcal{B}(1/\sqrt{\epsilon})$, it holds that $\mathcal{R}^{\circ}(w) \ge \mathcal{R}(w) - \epsilon_\ell$.*

Proof. Note that if $\ell_{\log}\left(y\langle w,x\rangle\right) \ne \ell^{\circ}_{\log}\left(y\langle w,x\rangle\right)$, then $y\langle w,x\rangle \le -B/\sqrt{\epsilon}$, which implies $\left|\langle w,x\rangle\right| \ge B/\sqrt{\epsilon}$. Moreover, in this case

$$\ell_{\log}\left(y\langle w,x\rangle\right) - \ell^{\circ}_{\log}\left(y\langle w,x\rangle\right) \le \ell_{\log}\left(y\langle w,x\rangle\right) - \ell_{\log}(0) \le \left|\langle w,x\rangle\right|.$$

Therefore

$$\mathcal{R}(w) - \mathcal{R}^{\circ}(w) = \mathbb{E}_{x\sim P_x}\left[\ell_{\log}\left(y\langle w,x\rangle\right) - \ell^{\circ}_{\log}\left(y\langle w,x\rangle\right)\right] \le \mathbb{E}_{x\sim P_x}\left[\left|\langle w,x\rangle\right| \mathbb{1}_{\left|\langle w,x\rangle\right|\ge B/\sqrt{\epsilon}}\right].$$

We can then invoke Equation (12) and get

$$\mathcal{R}(w) - \mathcal{R}^{\circ}(w) \le \alpha_1\left(\alpha_2\|w\| + \frac{B}{\sqrt{\epsilon}}\right)\exp\left(-\frac{B}{\alpha_2\|w\|\sqrt{\epsilon}}\right). \tag{21}$$

Note that the right hand side of Equation (21) is increasing with $\|w\|$, therefore we can let $\|w\|$ be $1/\sqrt{\epsilon}$ and get

$$\mathcal{R}(w) - \mathcal{R}^{\circ}(w) \le \alpha_1\frac{\alpha_2 + B}{\sqrt{\epsilon}}\exp\left(-\frac{B}{\alpha_2}\right) \le \epsilon_\ell,$$

where we use Equation (18) in the last inequality. □

Now putting everything together, under the conditions of Lemma C.11, with probability $1-2\delta$, projected gradient descent ensures $\mathcal{R}(w_t) \le \min_{0\le\rho\le 1/\sqrt{\epsilon}} \mathcal{R}(\rho\bar{u}) + 2\epsilon_\ell$. Moreover, by applying Lemma C.10 to Lemma C.11, we can see the sample complexity is $\widetilde{O}\left(d\ln(1/\delta)^3/(\epsilon\epsilon_\ell^2)\right)$, and the iteration complexity is $\widetilde{O}\left(d\ln(1/\delta)^2/(\epsilon\epsilon_\ell)\right)$.

C.2. Omitted proofs from Section 3.2

In this section, we prove Lemma C.4. We first prove the following approximation bound after we replace the true label with the label given by the ground-truth solution, which covers Lemma 3.11 and sub-exponential distributions.

Lemma C.13 (Lemma 3.11, including the sub-exponential case). *For $\ell \in \{\ell_{\log}, \ell_h\}$, if $\|x\| \le B$ almost surely,*

$$|\text{term (3)}| \le B\|\bar{w} - \hat{w}\| \cdot \mathrm{OPT}.$$

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

$$|\text{term (3)}| \le (1 + 2\alpha_1)\alpha_2\|\bar{w} - \hat{w}\| \cdot \mathrm{OPT} \cdot \ln(1/\mathrm{OPT}).$$

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

$$\text{term (3)} = \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y \ne \mathrm{sign}\left(\langle \bar{w}, x\rangle\right)} \cdot y\langle \bar{w} - \hat{w}, x\rangle\right], \tag{22}$$

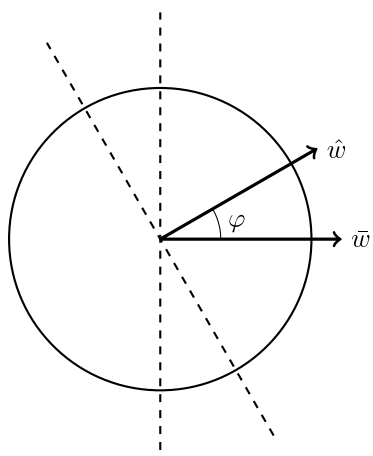
It then follows from the triangle inequality that

$$|\text{term (3)}| \le \mathbb{E}_{(x,y)\sim P}\left[\mathbb{1}_{y \ne \mathrm{sign}\left(\langle \bar{w}, x\rangle\right)}\left|\langle \bar{w} - \hat{w}, x\rangle\right|\right]$$

Now we can invoke Lemma A.2 with $w = \bar{w}$ and $w' = \bar{w} - \hat{w}$ to prove Lemma C.13. □

Next we prove the lower bound on term (4).

Proof of Lemma 3.12. Note that in term (4), 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 (4) is 0 on parts (ii) and (iv), therefore we only need to consider parts (i) and (iii):

$$\begin{aligned}\text{term (4)} &= \mathbb{E}_{\text{(i) and (iii)}}\left[\ell\left(\mathrm{sign}\left(\langle \bar{w}, x\rangle\right)\langle \hat{w}, x\rangle\right) - \ell\left(\mathrm{sign}\left(\langle \hat{w}, x\rangle\right)\langle \hat{w}, x\rangle\right)\right] \\ &= \mathbb{E}_{\text{(i) and (iii)}}\left[-\mathrm{sign}\left(\langle \bar{w}, x\rangle\right)\langle \hat{w}, x\rangle\right].\end{aligned}$$

Here we use the fact that $\ell(-z) - \ell(z) = z$ for both the logistic loss and the hinge 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 3.2, we have

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

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

Next, we prove the following upper bound on term (5), covering Lemma 3.13 and the sub-exponential case.

Lemma C.14 (Lemma 3.13, including the sub-exponential case). *For $\ell = \ell_h$, term (5) is 0. For $\ell = \ell_{\log}$, under Assumption 3.3, if $\|x\| \leq B$ almost surely, then*

$$|\text{term (5)}| \leq 12C_\kappa \cdot \frac{\varphi(\hat{w},\bar{w})}{\|\hat{w}\|},$$

where $C_\kappa := \int_0^B \kappa(r)\,\mathrm{d}r$, while if P_x is (α_1,α_2)-sub-exponential, then

$$|\text{term (5)}| \leq 2\alpha_1\mathrm{OPT}^2 + 12C_\kappa \cdot \frac{\varphi(\hat{w},\bar{w})}{\|\hat{w}\|},$$

where $C_\kappa := \int_0^{3\alpha_2\ln(1/\mathrm{OPT})} \kappa(r)\,\mathrm{d}r$.

Proof. For the hinge loss, term (5) is 0 simply because $\ell_h(z) = 0$ when $z \geq 0$. Next we consider the logistic loss.

Note that term (5) only depends on $\langle \hat{w}, x\rangle$ and $\langle \bar{w}, x\rangle$, therefore we can focus on the subspace spanned by $\hat{w}$ and $\bar{w}$. For simplicity, let p denote the density function of the projection of P_x onto the space spanned by $\hat{w}$ and $\bar{w}$. Moreover, without loss of generality we can assume $\bar{w}$ has polar angle 0 while $\hat{w}$ has polar angle φ, where we let φ denote $\varphi(\hat{w},\bar{w})$ for simplicity. It then follows that

$$\begin{aligned}
\text{term (5)} &= \int_0^\infty \int_0^{2\pi} \ell_{\log}\left(r\|\hat{w}\|\big|\cos(\theta-\varphi)\big|\right)p(r,\theta)r\,\mathrm{d}\theta\,\mathrm{d}r - \int_0^\infty \int_0^{2\pi} \ell_{\log}\left(r\|\hat{w}\|\big|\cos(\theta)\big|\right)p(r,\theta)r\,\mathrm{d}\theta\,\mathrm{d}r \\
&= \int_0^\infty \int_0^{2\pi} \ell_{\log}\left(r\|\hat{w}\|\big|\cos(\theta)\big|\right)\left(p(r,\theta+\varphi)-p(r,\theta)\right)r\,\mathrm{d}\theta\,\mathrm{d}r.
\end{aligned}$$

First, if $\|x\| \leq B$ almost surely, then

$$\begin{aligned}
|\text{term (5)}| &\leq \int_0^B \int_0^{2\pi} \ell_{\log}\left(r\|\hat{w}\|\big|\cos(\theta)\big|\right)\big|p(r,\theta+\varphi)-p(r,\theta)\big|\,r\,\mathrm{d}\theta\,\mathrm{d}r \\
&\leq \int_0^B \int_0^{2\pi} \ell_{\log}\left(r\|\hat{w}\|\big|\cos(\theta)\big|\right)\cdot\kappa(r)\varphi\cdot r\,\mathrm{d}\theta\,\mathrm{d}r \\
&= \varphi\int_0^B \kappa(r)\left(\int_0^{2\pi} \ell_{\log}\left(r\|\hat{w}\|\big|\cos(\theta)\big|\right)r\,\mathrm{d}\theta\right)\mathrm{d}r.
\end{aligned}$$

Then Lemma A.1 implies

$$|\text{term (5)}| \leq \varphi\int_0^B \kappa(r)\frac{8\sqrt{2}}{\|\hat{w}\|}\,\mathrm{d}r = 8\sqrt{2}C_\kappa\cdot\frac{\varphi}{\|\hat{w}\|} \leq 12C_\kappa\cdot\frac{\varphi}{\|\hat{w}\|}.$$

Next, assume P_x is (α_1, α_2)-sub-exponential. For a 2-dimensional random vector x sampled according to p_V, note that

$$\Pr\left(\|x\| \geq B\right) \leq \Pr\left(|x_1| \geq \frac{\sqrt{2}B}{2}\right) + \Pr\left(|x_2| \geq \frac{\sqrt{2}B}{2}\right) \leq 2\alpha_1 \exp\left(-\frac{\sqrt{2}B}{2\alpha_2}\right).$$

Letting $B := 2\sqrt{2}\alpha_2 \ln\left(\frac{1}{\mathrm{OPT}}\right)$, we get $\Pr\left(\|x\| \geq B\right) \leq 2\alpha_1 \mathrm{OPT}^2$. Since $\ell_{\log}(z) \leq 1$ when $z \geq 0$, we have

$$\begin{aligned}\text{term (5)} &\leq 2\alpha_1 \mathrm{OPT}^2 \\ &\quad + \int_0^B \int_0^{2\pi} \ell_{\log}\left(r\|\hat{w}\|\left|\cos(\theta-\varphi)\right|\right) p(r,\theta) r \,\mathrm{d}\theta \,\mathrm{d}r - \int_0^B \int_0^{2\pi} \ell_{\log}\left(r\|\hat{w}\|\left|\cos(\theta)\right|\right) p(r,\theta) r \,\mathrm{d}\theta \,\mathrm{d}r.\end{aligned}$$

Invoking the previous bound for bounded distributions, we get

$$\text{term (5)} \leq 2\alpha_1 \mathrm{OPT}^2 + 12 \cdot \frac{\varphi}{\|\hat{w}\|} \cdot \int_0^{2\sqrt{2}\alpha_2 \ln\left(\frac{1}{\mathrm{OPT}}\right)} \kappa(r) \,\mathrm{d}r \leq 2\alpha_1 \mathrm{OPT}^2 + 12 C_\kappa \cdot \frac{\varphi}{\|\hat{w}\|},$$

where $C_\kappa := \int_0^{3\alpha_2 \ln\left(\frac{1}{\mathrm{OPT}}\right)} \kappa(r) \,\mathrm{d}r$. Similarly, we can show

$$-\text{term (5)} \leq 2\alpha_1 \mathrm{OPT}^2 + 12 C_\kappa \cdot \frac{\varphi}{\|\hat{w}\|}.$$

□

Next we prove Lemma 3.14, which is basically (Diakonikolas et al., 2020d, Claim 3.4).

Proof of Lemma 3.14. Under Assumption 3.2, we have

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

□

Lastly, we prove Lemma C.4 for sub-exponential distributions.

Proof of Lemma C.4, sub-exponential distributions. For simplicity, let φ denotes $\varphi(\hat{w}, \bar{w})$. Lemmas 3.12, C.13 and C.14 imply

$$\begin{aligned}C_1 \|\hat{w}\| \varphi^2 &\leq \epsilon_\ell + C_2 \|\bar{w} - \hat{w}\| \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right) + C_3 \mathrm{OPT}^2 + C_4 C_\kappa \cdot \frac{\varphi}{\|\hat{w}\|} \\ &\leq \epsilon_\ell + C_2 \|\hat{w}\| \varphi \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right) + C_3 \mathrm{OPT}^2 + C_4 C_\kappa \cdot \frac{\varphi}{\|\hat{w}\|},\end{aligned}$$

where $C_1 = \frac{4R^3}{3U\pi^2}$, and $C_2 = (1 + 2\alpha_1)\alpha_2$, and $C_3 = 2\alpha_1$, and $C_4 = 12$. It follows that at least one of the following four cases is true:

1. $C_1 \|\hat{w}\| \varphi^2 \leq 4\epsilon_\ell$, which implies $\varphi = O\left(\sqrt{\epsilon_\ell / \|\hat{w}\|}\right)$.
2. $C_1 \|\hat{w}\| \varphi^2 \leq 4C_2 \|\hat{w}\| \varphi \cdot \mathrm{OPT} \cdot \ln\left(\frac{1}{\mathrm{OPT}}\right)$, which implies $\varphi = O\left(\mathrm{OPT} \ln\left(\frac{1}{\mathrm{OPT}}\right)\right)$.
3. $C_1 \|\hat{w}\| \varphi^2 \leq 4C_3 \mathrm{OPT}^2$, which implies $\varphi = O(\mathrm{OPT})$ since $\|\hat{w}\| = \Omega(1)$.
4. Lastly,

$$C_1 \|\hat{w}\| \varphi^2 \leq 4C_2 C_\kappa \cdot \frac{\varphi}{\|\hat{w}\|}, \quad \text{which implies} \quad \varphi = O\left(\frac{C_\kappa}{\|\hat{w}\|^2}\right). \tag{23}$$

Finally, we just need to invoke Lemma 3.14 to finish the proof. □

C.3. Omitted proofs from Section 3.3

We first prove the upper bound of term (5) under Assumption 3.2, without assuming the radially Lipschitz condition.

Proof of Lemma 3.15. Note that

$$\text{term (5)} \le \mathbb{E}\left[\ell_{\log}\left(\text{sign}\left(\langle \hat{w}, x\rangle\right)\langle \hat{w}, x\rangle\right)\right] = \mathbb{E}\left[\ell_{\log}\left(\left|\langle \hat{w}, x\rangle\right|\right)\right] \le \frac{12U}{\|\hat{w}\|},$$

where we invoke Lemma C.1 at the end. Similarly, we can show

$$-\text{term (5)} \le \frac{12U}{\|\bar{w}\|} = \frac{12U}{\|\hat{w}\|}$$

□

Next we prove a general result similar to Lemma C.4.

Theorem C.15. *Under Assumption 3.2, suppose* $\hat{w}$ *satisfies* $\mathcal{R}_{\log}(\hat{w}) \le \mathcal{R}_{\log}(\|\hat{w}\|\bar{u}) + \epsilon_\ell$ *for some* $\epsilon_\ell \in [0, 1)$*. If* $\|x\| \le B$ *almost surely, then*

$$\varphi(\hat{w}, \bar{u}) = O\left(\max\left\{\text{OPT}, \sqrt{\frac{\epsilon_\ell}{\|\hat{w}\|}}, \frac{1}{\|\hat{w}\|}\right\}\right).$$

If P_x *is* (α_1, α_2)*-sub-exponential and* $\|\hat{w}\| = \Omega(1)$*, then*

$$\varphi(\hat{w}, \bar{u}) = O\left(\max\left\{\text{OPT} \cdot \ln\left(\frac{1}{\text{OPT}}\right), \sqrt{\frac{\epsilon_\ell}{\|\hat{w}\|}}, \frac{1}{\|\hat{w}\|}\right\}\right).$$

Proof. For simplicity, let φ denote $\varphi(\hat{w}, \bar{u})$. Consider the case $\|x\| \le B$ almost surely. The condition $\mathcal{R}_{\log}(\hat{w}) \le \mathcal{R}_{\log}(\|\hat{w}\|\bar{u}) + \epsilon_\ell$, and Lemmas 3.12, 3.15 and C.13 imply

$$\begin{aligned} C_1\|\hat{w}\|\varphi^2 &\le \epsilon_\ell + B\|\bar{w} - \hat{w}\| \cdot \text{OPT} + \frac{C_2}{\|\hat{w}\|} \\ &\le \epsilon_\ell + B\|\hat{w}\|\varphi \cdot \text{OPT} + \frac{C_2}{\|\hat{w}\|}, \end{aligned}$$

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

1. $C_1\|\hat{w}\|\varphi^2 \le 3\epsilon_\ell$, which implies $\varphi = O\left(\sqrt{\epsilon_\ell/\|\hat{w}\|}\right)$;
2. $C_1\|\hat{w}\|\varphi^2 \le 3B\|\hat{w}\|\varphi \cdot \text{OPT}$, which implies $\varphi = O(\text{OPT})$;
3. $C_1\|\hat{w}\|\varphi^2 \le 3C_1/\|\hat{w}\|$, which implies $\varphi = O(1/\|\hat{w}\|)$.

The proof of the sub-exponential case is similar. □

Now we prove Lemma 3.16.

Proof of Lemma 3.16. First, if ϵ or OPT does not satisfy the conditions of Lemma C.7, then Lemma 3.16 holds vacuously; therefore in the following we consider the settings of Lemmas C.6 and C.7 with $\epsilon_\ell = \sqrt{\epsilon}$.

First, if $\|x\| \le B$ almost surely, Equation (15) and Theorem C.15 imply

$$\varphi(w_t, \bar{u}) = O\left(\max\left\{\text{OPT}, \sqrt{\frac{\epsilon_\ell}{\|w_t\|}}, \frac{1}{\|w_t\|}\right\}\right),$$

and moreover Lemma C.7 implies

$$\|w_t\| = \Omega\left(\min\left\{\frac{1}{\sqrt{\epsilon}}, \frac{1}{\sqrt{\mathrm{OPT}}}\right\}\right).$$

If $\epsilon \le \mathrm{OPT}$, then $\|w_t\| = \Omega\left(\frac{1}{\sqrt{\mathrm{OPT}}}\right)$, and

$$\begin{aligned}
\varphi(w_t, \bar{u}) &= O\left(\max\left\{\mathrm{OPT}, \sqrt{\epsilon_\ell\sqrt{\mathrm{OPT}}}, \sqrt{\mathrm{OPT}}\right\}\right) \\
&= O\left(\max\left\{\mathrm{OPT}, \sqrt{\sqrt{\epsilon}\sqrt{\mathrm{OPT}}}, \sqrt{\mathrm{OPT}}\right\}\right) \\
&= O\left(\max\left\{\mathrm{OPT}, \sqrt{\sqrt{\mathrm{OPT}}\sqrt{\mathrm{OPT}}}, \sqrt{\mathrm{OPT}}\right\}\right) = O\left(\sqrt{\mathrm{OPT}}\right).
\end{aligned}$$

If $\epsilon \ge \mathrm{OPT}$, then $\|w_t\| = \Omega\left(\frac{1}{\sqrt{\epsilon}}\right)$, and

$$\begin{aligned}
\varphi(w_t, \bar{u}) &= O\left(\max\left\{\mathrm{OPT}, \sqrt{\epsilon_\ell\sqrt{\epsilon}}, \sqrt{\epsilon}\right\}\right) \\
&= O\left(\max\left\{\mathrm{OPT}, \sqrt{\sqrt{\epsilon}\sqrt{\epsilon}}, \sqrt{\epsilon}\right\}\right) \\
&= O\left(\sqrt{\mathrm{OPT} + \epsilon}\right).
\end{aligned}$$

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

D. Omitted proofs from Section 4

In this section, we prove Theorem 4.1. We first prove a bound on $\mathcal{R}_h(\bar{u})$.

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

Proof. Note that

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

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

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

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

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

□

Next we prove the following result, which covers Lemma 4.2 but also handles sub-exponential distributions.

Lemma D.2 (Lemma 4.2, including the sub-exponential case). *Suppose Assumption 3.2 holds. Consider an arbitrary $w \in \mathcal{D}$, and let φ denote $\varphi(w, \bar{u})$. If $\|x\| \le B$ almost surely, then*

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

and

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

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

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

and

$$\begin{aligned}\mathcal{R}_h(w) - \mathcal{R}_h(\|w\|\bar{u}) \geq{}& \frac{4R^3}{3U\pi^2}\|w\|\varphi^2 \\ &- (1+2\alpha_1)\alpha_2\|w\|\varphi \cdot \mathrm{OPT} \cdot \ln(1/\mathrm{OPT}).\end{aligned}$$

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

$$\mathcal{R}_h(\bar{r}\bar{u}) = \frac{\bar{r}}{\|w\|}\mathcal{R}_h(\|w\|\bar{u}) \leq \mathcal{R}_h(\|w\|\bar{u}).$$

If $\bar{r} \geq \|w\|$, then

$$\mathcal{R}_h(\bar{r}\bar{u}) = \mathcal{R}_h\left(\|w\|\bar{u}\right) + \mathcal{R}_h(\bar{u})\left(\bar{r} - \|w\|\right) \leq \mathcal{R}_h\left(\|w\|\bar{u}\right) + \mathcal{R}_h(\bar{u})\left(\bar{r} - 1\right),$$

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

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

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,

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

where we apply Lemma D.1 at the end. It also follows directly from Lemmas 3.12, C.13 and C.14 that

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

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

Next we prove Theorem 4.1. We first consider the bounded case.

Proof of Theorem 4.1, bounded distribution. Here we assume $\|x\| \leq B$ almost surely. We will show that under the conditions of Theorem 4.1, then

$$\mathbb{E}\left[\min_{0 \leq t < T} \varphi_t\right] = O(\mathrm{OPT} + \epsilon), \quad \text{where} \quad \varphi_t := \varphi(w_t, \bar{u}). \tag{24}$$

Further invoking Lemma 3.14 finishes the proof.

Recall that at step t, after taking the expectation with respect to (x_t, y_t), we have

$$\begin{aligned}\mathbb{E}\left[\|w_{t+1} - \bar{r}\bar{u}\|^2\right] &\le \|w_t - \bar{r}\bar{u}\|^2 - 2\eta\left\langle \nabla\mathcal{R}_h(w_t), w_t - \bar{r}\bar{u}\right\rangle + \eta^2 B^2 \mathcal{M}(w_t) \\ &\le \|w_t - \bar{r}\bar{u}\|^2 - 2\eta\left(\mathcal{R}_h(w_t) - \mathcal{R}_h(\bar{r}\bar{u})\right) + \eta^2 B^2 \mathcal{M}(w_t). \end{aligned} \tag{25}$$

First, Lemma D.2 implies

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

where $C_1 := 2R^3/(3U\pi^2)$. Note that if $\varphi_t \le B \cdot \mathrm{OPT}/C_1$, then Equation (24) holds; therefore in the following we assume

$$\varphi_t \ge \frac{B}{C_1} \cdot \mathrm{OPT}, \tag{26}$$

which implies

$$\mathcal{R}_h(w_t) - \mathcal{R}_h(\bar{r}\bar{u}) \ge C_1\|w_t\|\varphi_t^2 - O\left((\mathrm{OPT} + \epsilon)^2\right) \ge C_1\varphi_t^2 - O\left((\mathrm{OPT} + \epsilon)^2\right), \tag{27}$$

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

On the other hand, Equation (26) and Lemma 3.14 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 Equation (24) is true; otherwise we can assume $\epsilon \le \varphi_t$, and let $\eta = C_2\epsilon$, we have

$$\eta B^2 \mathcal{M}(w_t) \le C_2\epsilon B^2\left(\frac{C_1}{B} + 2U\right)\varphi_t = C_1\epsilon\varphi_t \le C_1\varphi_t^2. \tag{28}$$

Now Equations (25), (27) and (28) imply

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

Taking the expectation and average, we have

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

Note that

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

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 Equation (24) holds. □

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

Lemma D.3. *Suppose P_x is (α_1, α_2)-sub-exponential. Given any threshold $\tau > 0$, it holds that*

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

Proof. First recall that

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

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

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

Calculation gives

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

□

Now we are ready to prove Theorem 4.1 for sub-exponential distributions.

Proof of Theorem 4.1, sub-exponential distributions. At step t, we have

$$\begin{aligned}\|w_{t+1} - \bar{r}\bar{u}\|^2 &\leq \|w_t - \bar{r}\bar{u}\|^2 - 2\eta\left\langle \ell_h'\left(y_t\langle w_t, x_t\rangle\right) y_t x_t, w_t - \bar{r}\bar{u}\right\rangle + \eta^2 \ell_h'\left(y_t\langle w_t, x_t\rangle\right)^2 \|x_t\|^2 \\ &= \|w_t - \bar{r}\bar{u}\|^2 - 2\eta\left\langle \ell_h'\left(y_t\langle w_t, x_t\rangle\right) y_t x_t, w_t - \bar{r}\bar{u}\right\rangle - \eta^2 \ell_h'\left(y_t\langle w_t, x_t\rangle\right) \|x_t\|^2, \end{aligned} \tag{29}$$

where we use $(\ell_h')^2 = -\ell_h'$. Next we bound $\mathbb{E}_{(x_t,y_t)}\left[-\ell_h'\left(y_t\langle w_t, x_t\rangle\right)\|x_t\|^2\right]$. Let $\tau := \sqrt{d}\alpha_2\ln(d/\epsilon)$. When $\|x_t\| \leq \tau$, we have

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

On the other hand, when $\|x_t\| \geq \tau$, Lemma D.3 implies

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

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

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

for some constant C.

Now taking the expectation with respect to (x_t, y_t) on both sides of Equation (29), we have

$$\mathbb{E}\left[\|w_{t+1} - \bar{r}\bar{u}\|^2\right] \leq \|w_t - \bar{r}\bar{u}\|^2 - 2\eta\left(\mathcal{R}_h(w_t) - \mathcal{R}_h(\bar{r}\bar{u})\right) + \eta^2 Cd\left(\mathcal{M}(w_t) + \epsilon\right)\cdot\ln(d/\epsilon)^2. \tag{30}$$

Similarly to the bounded case, we will show that

$$\mathbb{E}\left[\min_{0 \leq t < T}\varphi_t\right] = O\left(\mathrm{OPT}\cdot\ln(1/\mathrm{OPT}) + \epsilon\right), \quad \text{where} \quad \varphi_t := \varphi(w_t, \bar{u}). \tag{31}$$

First, Lemma D.2 implies

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

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 Equation (31) holds; therefore in the following we assume

$$\varphi_t \geq \frac{C_2}{C_1} \cdot \mathrm{OPT} \cdot \ln(1/\mathrm{OPT}), \tag{32}$$

which implies

$$\begin{aligned}\mathcal{R}_h(w_t) - \mathcal{R}_h(\bar{r}\bar{u}) &\geq C_1\|w_t\|\varphi_t^2 - O\left(\left(\mathrm{OPT} \cdot (1/\mathrm{OPT}) + \epsilon\right)^2\right) \\ &\geq C_1\varphi_t^2 - O\left(\left(\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon\right)^2\right), \end{aligned}\tag{33}$$

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

On the other hand, for $\mathrm{OPT} \leq 1/e$, Equation (32) and Lemma 3.14 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 Equation (31) is true; otherwise we can assume $\epsilon \leq \varphi_t$, and let $\eta = \frac{C_2\epsilon}{d\ln(d/\epsilon)^2}$, we have

$$\begin{aligned}\eta C d\left(\mathcal{M}(w_t) + \epsilon\right)\ln(d/\epsilon)^2 &= \frac{C_2\epsilon}{d\ln(d/\epsilon)^2}Cd\mathcal{M}(w_t) \cdot \ln(d/\epsilon)^2 + \frac{C_2\epsilon}{d\ln(d/\epsilon)^2}Cd\epsilon \cdot \ln(d/\epsilon)^2 \\ &\leq C_2\epsilon C\left(\frac{C_1}{C_2} + 2U\right)\varphi_t + C_2C\epsilon^2 \\ &= C_1\epsilon\varphi_t + O\left(\left(\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon\right)^2\right) \\ &\leq C_1\varphi_t^2 + O\left(\left(\mathrm{OPT} \cdot \ln(1/\mathrm{OPT}) + \epsilon\right)^2\right). \end{aligned}\tag{34}$$

Now Equations (30), (33) and (34) imply

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

Taking the expectation and average, we have

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

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

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

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

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

and thus Equation (31) holds. □