Learning General Halfspaces with Adversarial Label Noise via Online Gradient Descent

Ilias Diakonikolas [*1] **Vasilis Kontonis** [*1] **Christos Tzamos** [*1] **Nikos Zarifis** [*1]

Abstract

We study the problem of learning general — i.e., not necessarily homogeneous — halfspaces with adversarial label noise under the Gaussian distribution. Prior work has provided a sophisticated polynomial-time algorithm for this problem. In this work, we show that the problem can be solved directly via online gradient descent applied to a sequence of natural non-convex surrogates. This approach yields a simple iterative learning algorithm for general halfspaces with near-optimal sample complexity, runtime, and error guarantee. At the conceptual level, our work establishes an intriguing connection between learning halfspaces with adversarial noise and online optimization that may find other applications.

1. Introduction

We study the distribution-specific PAC learnability of linear classifiers (or halfspaces) in the presence of adversarial label noise. Before we describe our contributions, we provide the necessary context for this work.

1.1. Background and Motivation

A *halfspace* or Linear Threshold Function (LTF) is any Boolean-valued function $f : \mathbb{R}^d \mapsto \{\pm 1\}$ of the form $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t)$, for a vector $\mathbf{w}^* \in \mathbb{R}^d$ (known as the weight vector) and a scalar $t \in \mathbb{R}$ (known as the threshold). Halfspaces are a central class of Boolean functions that arise in several areas of computer science, including complexity theory, learning theory, and optimization (Rosenblatt, 1958; Novikoff, 1962; Minsky & Papert, 1968; Yao, 1990; Goldmann et al., 1992; Freund & Schapire, 1997; Vapnik, 1998; Shawe-Taylor & Cristianini, 2000; O'Donnell, 2014). Here we focus on the algorithmic problem of learning halfspaces from labeled examples, arguably one of *the* most extensively studied and influential problems in machine learning.

The computational problem of learning halfspaces from random examples is efficiently solvable without noise (Maass & Turan, 1994) in the distribution-independent setting. The complexity of the problem in the presence of corrupted data depends on the contamination model and the underlying distributional assumptions. In this paper, we focus on learning with adversarial label noise under the Gaussian distribution. Formally, we have the following definition.

Definition 1.1 (Learning with Adversarial Label Noise)**.** Let $\mathcal{C}$ be a concept class of Boolean functions over $X = \mathbb{R}^d$ and $\epsilon \in (0, 1/2)$. Let f be an unknown target function in $\mathcal{C}$. A *noisy example oracle*, $\mathrm{EX}(f, \epsilon)$, works as follows: Each time $\mathrm{EX}(f, \epsilon)$ is invoked, it returns a labeled example $(\mathbf{x}, y)$, such that: (a) $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$, where $\mathcal{D}_{\mathbf{x}}$ is the standard normal distribution on $\mathbb{R}^d$, and (b) $y \neq f(\mathbf{x})$ with probability at most ϵ. Let $\mathcal{D}$ denote the joint distribution on $(\mathbf{x}, y)$ generated by the above oracle. We say that such a distribution $\mathcal{D}$ is *ϵ-corrupted*. For some constant $C \geq 1$, a C-approximate learning algorithm is given i.i.d. samples from $\mathcal{D}$ and its goal is to output a hypothesis h such that with high probability it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq C\epsilon$.

We remark that the agnostic PAC-learning model (Kearns et al., 1994) requires that the hypothesis $h(\mathbf{x})$ satisfies $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y] + \epsilon'$, for any desired $\epsilon' > 0$. This guarantee corresponds to the $C = 1$ case of Definition 1.1. Without distributional assumptions, learning with adversarial label noise is computationally hard for any constant $C > 1$ (Daniely, 2016). This motivates studying the problem under natural distributional assumptions. Interestingly, it is known (Klivans & Kothari, 2014) that even when the underlying distribution is the standard Gaussian, achieving an agnostic error guarantee (i.e., $C = 1$ in Definition 1.1) requires super-polynomial runtime (under a plausible complexity assumption about the hardness of noisy parity). The latter hardness result does not rule out constant factor approximations, i.e., achieving approximation ratio $C > 1$, where C is some *universal constant*. Indeed, this is the regime where the algorithmic results of this paper apply.

[*]Equal contribution [1]Department of Computer Sciences, University of Wisconsin, Madison, Wisconsin, USA. Correspondence to: Vasilis Kontonis <kontonis@wisc.edu>, Nikos Zarifis <zarifis@wisc.edu>.

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

General vs Homogeneous Halfspaces A halfspace is called *homogeneous* if the defining separating hyperplane goes through the origin; or equivalently if it can be expressed in the form $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. The vast majority of prior work on efficiently learning halfspaces with adversarial label noise (under natural distributional assumptions) is restricted the homogeneous case. This line of work was initiated in (Awasthi et al., 2017) who gave the first polynomial time constant-factor approximate learner under isotropic log-concave distributions. Subsequently, a number of works generalized and/or quantitatively improved on that work (Diakonikolas et al., 2018; 2020c; Shen, 2021).

Interestingly, with the sole exception of (Diakonikolas et al., 2018), all prior approaches for our problem inherently fail when the halfspace is no longer homogeneous, i.e., when a non-zero threshold t is introduced. Moving from homogeneous to general halfspaces may seem innocuous at first sight. Indeed, it is seemingly straightforward to reduce a general halfspace to a homogeneous one by adding an extra constant coordinate to every sample. While this reduction is valid in the distribution-independent setting, it does not work in the distribution-specific setting because it alters the marginal distribution on the examples. In fact, for general halfspaces, the only known result that achieves error $O(\epsilon)$ in polynomial time is from (Diakonikolas et al., 2018). (We note that their algorithm also succeeds in a more general contamination model that also allows an ϵ-fraction of the points to be corrupted in addition to the labels.)

Motivation The learning algorithm of (Diakonikolas et al., 2018) for general halfspaces has some shortcomings. First, the algorithm is rather complicated and consequently seems difficult to generalize to other settings. Second, its sample complexity and runtime, while polynomially bounded, are significantly sub-optimal. Ideally, we would like a simple and practical iterative algorithm for the problem that moreover has (near-)optimal sample size and runtime.

Concretely, we ask whether we can achieve the near-optimal error guarantee of (Diakonikolas et al., 2018) with a simple and practical algorithm.

Is there a simple iterative method to learn general halfspaces with an ϵ-fraction of adversarial label noise?

Our main result provides an affirmative answer to this question. Specifically, we develop a learning algorithm based on online gradient descent that additionally achieves near-optimal sample complexity and runtime.

1.2. Our Results and Techniques

Our main result is a simple gradient-based iteration that efficiently converges to a halfspace with error $O(\epsilon)$ after only $\mathrm{poly}\log(1/\epsilon)$ rounds. We show the following theorem (see Theorem 4.1 for the formal statement).

Theorem 1.2 (Online Gradient Descent Learner)**.** *Let $\epsilon \in (0, 1/2)$ and $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Using $N = \widetilde{O}((d/\epsilon^2)\log(1/\delta))$ samples, we can construct a sequence of $T = \mathrm{poly}(\log(1/\epsilon))$ non-convex loss-functions such that Online Projected-Gradient-Descent converges to a vector $\mathbf{w}^{(T)}$ that, with probability at least $1 - \delta$, satisfies $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(\mathbf{w}^{(T)} \cdot \mathbf{x} + t) \neq y] \leq C\epsilon$, for a universal constant C. The resulting algorithm has runtime $\widetilde{O}(N)\, d$.*

The connection with online non-convex optimization drawn in Theorem 1.2 is new and may find other applications. In addition to obtaining a much simpler algorithm, Theorem 1.2 leads to the first sample (and time) near-optimal algorithm for the problem. Indeed, the algorithm given in the prior work (Diakonikolas et al., 2018) has sample complexity and runtime $\mathrm{poly}(d/\epsilon)$, for some polynomial of unspecified degree.

Discussion Our algorithmic result is enabled by a novel non-convex feasibility program that essentially characterizes approximate learnability of halfspaces with adversarial label noise under the Gaussian distribution. Our non-convex formulation leverages the idea of localization, i.e., focusing on the samples that fall in some specific subset of the space. Localization is a powerful tool going back to the work of (Bartlett et al., 2005). Variants of the method have been used in several works that deal with learning noisy halfspaces (see, e.g., (Awasthi et al., 2015; 2016; Yan & Zhang, 2017; Zhang et al., 2020; Diakonikolas et al., 2020c)). While localization is most commonly used to "zoom in" on the region close to decision boundary of the classifier, in this work we will be focusing on arbitrary subsets that can be far from the decision boundary. Specifically, we consider intervals of the form $B_{a,b} = \{z \in \mathbb{R} : |z - b| \leq a\}$; and for a given a weight vector $\mathbf{w}$, we restrict our attention to the band $\mathbf{w} \cdot \mathbf{x} \in B_{a,b}$. Our main insight is that, in the noiseless setting, i.e., when $y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t)$, the label y is constant along any direction orthogonal to the optimal direction $\mathbf{w}^*$, and therefore for any $a \geq 0, b \in \mathbb{R}$ it holds $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{(\mathbf{w}^*)^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}^* \cdot \mathbf{x} \in B_{a,b}\}] = \mathbf{0}$, where $\mathrm{proj}_{\mathbf{u}^\perp}\mathbf{x}$ is the projection of $\mathbf{x}$ onto the orthogonal complement of $\mathbf{u}$. On the other hand, for any direction $\mathbf{w}$ not parallel to $\mathbf{w}^*$, it is not hard to see that $\|\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B_{a,b}\}]\|_2 > 0$, since $\mathbf{w}^*$ has now non-zero projection onto the subspace $\mathbf{w}^\perp$. When there is noise in the data, it may happen that for many bands $B_{a,b}$ it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B_{a,b}\}] = \mathbf{0}$ even though the weight vector $\mathbf{w}$ is far from the optimal $\mathbf{w}^*$. However, we show that since only an ϵ-fraction of the examples are noisy, this cannot happen for all bands, and, in particular, when $\mathbf{w}$ is far from being optimal, there must exist some band inside

which $\|\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{(\mathbf{w}^*)^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\in B_{a,b}\}]\|_2$, is significantly smaller than $\|\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}]\|_2$. The key property is that there exists a carefully designed threshold on the value of the norm $\|\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}]\|_2$ that makes the optimal weight-vector feasible and any significantly sub-optimal vector $\mathbf{w}$ infeasible. The following feasibility problem [1] formalizes this idea:

$$\begin{aligned}\text{Find}\quad & \mathbf{w}:\|\mathbf{w}\|_2=1\\ \text{s. t.}\quad & \left\|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\left\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\right\}]\right\|_2\le\\ & 4\sqrt{e}\,\epsilon\sqrt{\log\left(\mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[z\in B_{a,b}]/\epsilon+1\right)} \qquad (1)\\ & \forall a\ge 0, b\in\mathbb{R}\end{aligned}$$

Our main structural result shows that the above non-convex feasibility program essentially identifies an approximately optimal weight vector up to a sign change.

Theorem 1.3 (Non-Convex Feasibility Program)**.** *Fix $\epsilon\in(0,1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d\times\{\pm1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\mathbf{w}^*$ the weight vector of an optimal halfspace, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t)\neq y]\le\epsilon$. We have that: (i) program (1) is feasible, and (ii) any solution $\mathbf{w}$ satisfies $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq y]\le C\epsilon$ or $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(-\mathbf{w}\cdot\mathbf{x}+t)\neq y]\le C\epsilon$, where C is some universal constant.*

Non-convex problems are of course computationally intractable in general. Therefore, at first sight, Theorem 1.3 cannot readily be used to obtain an efficient algorithm. Our algorithmic result relies on the following crucial property: given any sub-optimal weight vector $\mathbf{w}$, we can identify a band $B_{a,b}$ that corresponds to a violated constraint of the feasibility problem (1). We then use this band to improve the guess $\mathbf{w}$. Our algorithm fits in the online optimization framework: at every round the (optimization) adversary tries to find a band $B_{a,b}$ that corresponds to a violated constraint of (1) and then presents the learner with a loss-function that forces them to focus on the band $B_{a,b}$. We show that Online Gradient Descent with these "localized" loss-functions converges to an almost optimal solution.

1.3. Related Work

In the preceding discussion, we have already mentioned the most closely related prior work. Here we elaborate on some of these results. The work (Diakonikolas et al., 2020c) shows that SGD on a carefully selected non-convex surrogate achieves error $O(\epsilon)$ for the special case of homogeneous halfspaces. Unfortunately, this method fails for general halfspaces, and it is not clear whether an alternative formulation exists. In the same vein, (Shen, 2021) gave a (localized) perceptron update rule to learn halfspaces with adversarial noise under structured marginals. Our online convex optimization approach is more powerful and allows us to circumvent the obstacles posed when restricted to a single objective.

Finally, we acknowledge the work of (Frei et al., 2020) which developed gradient-based algorithms for learning ReLU activations under structured distributions. Similarly, (Frei et al., 2020) studies the homogeneous case, corresponding to functions of the form $\mathrm{ReLU}(\mathbf{w}^*\cdot\mathbf{x})$. For the case of Gaussian marginals (that is relevant to our work), the error guarantee they obtain is of the form $O(d\,\epsilon)$, where ϵ is the optimal L_2-loss.

2. Preliminaries

We use small boldface characters for vectors. For $\mathbf{x}\in\mathbb{R}^d$ and $i\in[d]$, $\mathbf{x}_i$ denotes the i-th coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2:=(\sum_{i=1}^d\mathbf{x}_i^2)^{1/2}$ denotes the ℓ_2-norm of $\mathbf{x}$. We will use $\mathbf{x}\cdot\mathbf{y}$ for the inner product of $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ and $\theta(\mathbf{x},\mathbf{y})$ for the angle between $\mathbf{x},\mathbf{y}$. For simplicity of notation, we may use θ instead of $\theta(\mathbf{x},\mathbf{y})$ when it is clear from the context. We will use $\mathbb{1}_A$ to denote the characteristic function of the set A, i.e., $\mathbb{1}_A(\mathbf{x})=1$ if $\mathbf{x}\in A$ and $\mathbb{1}_A(\mathbf{x})=0$ if $\mathbf{x}\notin A$. Let $\mathbf{e}_i$ be the i-th standard basis vector in $\mathbb{R}^d$. For a vector $\mathbf{w}\in\mathbb{R}^d$, we use $\mathbf{w}^\perp$ to denote the subspace spanned by vectors orthogonal to $\mathbf{w}$, i.e., $\mathbf{w}^\perp=\{\mathbf{u}\in\mathbb{R}^d:\mathbf{w}\cdot\mathbf{u}=0\}$. For a subspace $U\subseteq\mathbb{R}^d$, we denote $(\mathrm{proj}_U\mathbf{x})$, the projection of $\mathbf{x}$ onto U.

We use $\mathbf{E}_{x\sim\mathcal{D}}[x]$ for the expectation of the random variable x according to the distribution $\mathcal{D}$ and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. For simplicity of notation, we may omit the distribution when it is clear from the context. Let $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denote the d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}\in\mathbb{R}^d$ and covariance $\boldsymbol{\Sigma}\in\mathbb{R}^{d\times d}$. we also denote $\mathcal{N}(\mu,\sigma^2)$, the standard normal distribution with mean μ and variance σ^2. For $(\mathbf{x},y)$ distributed according to $\mathcal{D}$, we denote $\mathcal{D}_\mathbf{x}$ to be the distribution of $\mathbf{x}$ and $\mathcal{D}_y$ to be the distribution of y. For a set B and a distribution $\mathcal{D}$, we denote $\mathcal{D}_B$ to be the distribution $\mathcal{D}$ conditional on B. Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal, i.e., $\Phi(t)=1/\sqrt{2\pi}\int_{-\infty}^t\exp(-z^2/2)\mathrm{d}z$, moreover, we denote $\Phi^{-1}(\cdot)$ to be the inverse function of $\Phi(\cdot)$.

[1] Observe that the variable of the non-convex program is only the weight vector $\mathbf{w}$. We remark that we can always assume that we know the optimal threshold since it is straightforward to estimate it (even with noisy samples), see also Claim 4.7.

3. Structural Result: A Non-Convex Feasibility Problem

In this section, we prove our structural result and show that the solutions of the feasibility problem (1) are near optimal halfspaces. We remark that, in general, it is not hard to construct non-convex feasibility programs with optimal solutions: minimizing the zero-one loss is indeed such a non-convex problem which is known to be computationally challenging under adversarial label noise even when the underlying distribution is the standard normal. Our non-convex feasibility formulation is inherently different than the standard zero-one loss minimization and the proof of its identifiability is the basis of our Online Gradient Descent algorithm.

Before we prove the theorem, we would like to highlight its connection with our algorithmic result. Our Online Gradient Descent algorithm essentially uses as gradients the vectors in the left-hand side of the constraint of the non-convex program (1). In Section 4, we will show that, as long as the current halfspace h is not nearly optimal, we can find an appropriate band $B_{a,b}$ and use the vector $\mathbf{g} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}]$ to improve the guess $\mathbf{w}$. For the above reasons, we will refer to this vector $\mathbf{g}$ as the "gradient".

We split the proof of Theorem 1.3 in two parts: in Lemma 3.1, where we show that the non-convex program is feasible and Lemma 3.3, where we show that any solution is an approximately optimal halfspace.

Lemma 3.1 (Feasibility)**.** *Fix $\epsilon\in(0,1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d\times\{\pm1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\mathbf{w}^*$ the weight vector of an optimal halfspace, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t)\neq y]\leq\epsilon$. Then $\mathbf{w}^*$ is a feasible solution of the non-convex problem (1).*

Proof. Note that $\|\mathbf{g}\|_2=\sup_{\mathbf{v}\in\mathbb{R}^d}|\frac{\mathbf{v}}{\|\mathbf{v}\|_2}\cdot\mathbf{g}|$. Pick any unit vector $\mathbf{u}\in\mathbb{R}^d$ and denote $f(\mathbf{x})=\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t)$. Using the triangle inequality and the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f(\mathbf{x})\mathbf{u}\cdot(\mathrm{proj}_{(\mathbf{w}^*)^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\in B_{a,b}\}]=0$, we have that

$$\begin{aligned}
&|\mathbf{u}\cdot\mathbf{g}|\\
&\leq|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f(\mathbf{x})\mathbf{u}\cdot(\mathrm{proj}_{(\mathbf{w}^*)^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\in B_{a,b}\}]|\\
&+|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-f(\mathbf{x}))\mathbf{u}\cdot(\mathrm{proj}_{(\mathbf{w}^*)^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\in B_{a,b}\}]|\\
&\leq 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u}\cdot(\mathrm{proj}_{(\mathbf{w}^*)^\perp}\mathbf{x})|\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\in B_{a,b},y\neq f(\mathbf{x})\}]\,.
\end{aligned}$$

We now have to bound from above the contribution of the term $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u}\cdot(\mathrm{proj}_{(\mathbf{w}^*)^\perp}\mathbf{x})|\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\in B_{a,b},y\neq f(\mathbf{x})\}]$. One could use the Cauchy-Schwarz inequality to bound this expectation by $\sqrt{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u}\cdot(\mathrm{proj}_{(\mathbf{w}^*)^\perp}\mathbf{x})|^2]}\sqrt{\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{w}^*\cdot\mathbf{x}\in B_{a,b},y\neq f(\mathbf{x})]}$. However, this would only imply an upper bound of the order of $(\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{w}^*\cdot\mathbf{x}\in B_{a,b},y\neq f(\mathbf{x})])^{1/2}=O(\sqrt{\epsilon})$. Using the concentration of the Gaussian distribution and the fact that $\mathbf{u}\cdot\mathrm{proj}_{(\mathbf{w}^*)^\perp}\mathbf{x}$ is independent from $\mathbf{w}^*\cdot\mathbf{x}$ we are able to prove a much stronger decoupling inequality. We show the following lemma.

Lemma 3.2 (Gaussian Decoupling Inequality)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$ with standard normal $\mathbf{x}$-marginal. Moreover, let $\mathbf{w},\mathbf{u}\in\mathbb{R}^d$ be two orthogonal unit vectors, define $B=\{z\in\mathbb{R}:z\in(t_1,t_2)\}$, for some $t_1,t_2\in\mathbb{R}$ and let $S(\mathbf{x},y)$ be an event over $\mathbb{R}^d\times\{\pm1\}$. It holds that*

$$\begin{aligned}
&\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{S(\mathbf{x},y),\mathbf{w}\cdot\mathbf{x}\in B\}]\\
&\leq 2\sqrt{e}\,\mathbf{Pr}[S(\mathbf{x},y)]\sqrt{\log\left(\frac{\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x}\in B]}{\mathbf{Pr}[S(\mathbf{x},y)]}+1\right)}\,.
\end{aligned}$$

Using Lemma 3.2, we get that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b},y\neq h(\mathbf{x})\}]\\
&\leq 2\sqrt{e}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq y]\sqrt{\log\left(\frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z\in B_{a,b}]}{\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq y]}\right)}\\
&\leq 2\sqrt{e}\epsilon\sqrt{\log\left(\frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z\in B_{a,b}]}{\epsilon}\right)}\,.
\end{aligned}$$

□

Next we show that for any vector $\mathbf{w}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq y]\geq C\epsilon$, where $C>0$ is some sufficiently large universal constant, there exists a set B which violates a constraint of the non-convex program (1). In particular we show that by choosing $a=\sin(\theta(\mathbf{w},\mathbf{w}^*))$ and $b=-t\cos(\theta(\mathbf{w},\mathbf{w}^*))$ we get a violated constraint of problem (1). We establish the following lemma.

Lemma 3.3 (Approximate Robust Identifiability)**.** *Fix $\epsilon\in(0,1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d\times\{\pm1\}$ with standard normal $\mathbf{x}$-marginal. Let $\mathbf{w}\in\mathbb{R}^d$ be any feasible solution to Equation (1). It holds that either $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq y]\leq C\epsilon$ or $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(-\mathbf{w}\cdot\mathbf{x}+t)\neq y]\leq C\epsilon$, for some universal constant $C\geq1$.*

Remark 3.4. To prove the lemma, we consider any vector $\mathbf{w}$ with significantly sub-optimal error and then show that there exists a band $B_{a,b}$ such that the norm of the gradient $\|\mathbf{g}\|_2$ violates the corresponding constraint of problem (1). In fact, we prove a stronger statement that will eventually allow us to design an efficient algorithm. We show that the vector $\mathbf{g}=\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}]$ correlates positively with the optimal vector $\mathbf{w}^*$, i.e., that $\mathbf{g}\cdot\mathbf{w}^*>0$. At a high-level, this means that $\mathbf{g}$ "points to the right direction" and we can use it in order to improve our current guess; see Lemma 3.5 and Section 4.

Proof. As we plan to bound from below the norm of $\mathbf{g}$, it suffices to consider any unit vector $\mathbf{u}$ and show that $\mathbf{g} \cdot \mathbf{u}$ is large. We will first prove a general lemma that, given any band B, bounds from below the inner product of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})y\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B\}]$ with the optimal direction $\mathbf{w}^*$.

Lemma 3.5 (Noisy Gradient Decomposition)**.** *Fix $\epsilon \in (0, 1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\mathbf{w}^*$ with $\|\mathbf{w}^*\|_2 = 1$, the weight vector of an optimal halfspace, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t)\neq y] \leq \epsilon$. Fix some unit vector $\mathbf{w}\in\mathbb{R}^d$ such that $\theta(\mathbf{w},\mathbf{w}^*)=\theta\in(0,\pi)$ and let $B=\{z\in\mathbb{R}: t_1<z<t_2\}$. Denote $\mathbf{g}=\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B\}]$ and $\mathbf{v}=\frac{\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*}{\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2}$. It holds*

$$\mathbf{g}\cdot\mathbf{v} \geq \sqrt{\frac{2}{\pi}}\left(\sin\theta e^{-\frac{t^2}{2}}p - 9\epsilon\sqrt{\log(q/\epsilon+1)}\right),$$

where $p=\mathbf{Pr}_{z\sim\mathcal{N}(-t\cos\theta,(\sin\theta)^2)}[z\in B]$, and $q=\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z\in B]$.

Remark 3.6. The term $\sin\theta e^{-\frac{t^2}{2}}p$ in the lower bound of $\mathbf{g}\cdot\mathbf{v}$ in Lemma 3.5 corresponds to the contribution of the clean examples, i.e., when the noise rate is $\epsilon=0$ we would only get this term. The term $\epsilon\sqrt{\log(q/\epsilon+1)}$ corresponds to the noise. This can potentially make $\mathbf{g}\cdot\mathbf{v}<0$, and therefore make the gradient vector $\mathbf{g}$ point away from $\mathbf{w}^*$.

We are now ready to prove Lemma 3.3. We consider any $\mathbf{w}$ with both $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq y] > C\epsilon$ and $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(-\mathbf{w}\cdot\mathbf{x}+t)\neq y] > C\epsilon$. We will show that there exists $B_{a,b}$ such that $\|\mathbf{g}\|_2 > 4\sqrt{e}\sqrt{\log(\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[B_{a,b}]/\epsilon+1)}$, i.e., a constraint of the program (1) is violated. To bound from below the norm of $\|\mathbf{g}\|_2$ we can pick any unit vector $\mathbf{v}$ and obtain the lower bound $\|\mathbf{g}\|_2\geq\mathbf{g}\cdot\mathbf{v}$. We choose the vector $\mathbf{v}=\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*/\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2$ and use Lemma 3.5, to obtain

$$\|\mathbf{g}\|_2 \geq \sqrt{\frac{2}{\pi}}\left(\frac{2}{3}\sin\theta e^{-\frac{t^2}{2}} - 9\epsilon\sqrt{\log(q/\epsilon+1)}\right). \quad (2)$$

Our goal is to show that under the assumption that the error of the weight vector $\mathbf{w}$ is larger than $C\epsilon$ the positive term $\sin\theta e^{-t^2/2}$ of the above inequality is much larger than the "noise" contribution $\epsilon\sqrt{\log(q/\epsilon+1)}$.

Recall, that in Lemma 3.3 we pick the band $B_{a,b}$ with $a=\sin(\theta(\mathbf{w},\mathbf{w}^*))$ and $b=-t\cos(\theta(\mathbf{w},\mathbf{w}^*))$. We now have to provide estimates for the probabilities p and q of Lemma 3.5.

$$p = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(-t\cos\theta,(\sin\theta)^2)}[z\in B]$$
$$= \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,(\sin\theta)^2)}[|z|\leq\sin\theta] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[|z|\leq 1] \geq \frac{2}{3}.$$

Moreover, we have

$$q = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[z\in B] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[|z+t\cos\theta|\leq\sin\theta]$$
$$\leq 4\sin\theta e^{-t^2/2}\, e^{(t\sin\theta)^2/2+|t|\sin\theta}.$$

In order to bound from below the term $(2/3)\sin\theta e^{-\frac{t^2}{2}}$, we will use our assumption that the error of the current weight vector $\mathbf{w}$ is larger than $C\epsilon$. To do so, we need to connect the disagreement between two halfspaces with their parameter distance, i.e., the angle of their normal vectors. We will use the following lemma.

Lemma 3.7 ((Diakonikolas et al., 2018))**.** *For unit vectors $\mathbf{w}_1,\mathbf{w}_2\in\mathbb{R}^d$ and $t\in\mathbb{R}$, it holds*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{sign}(\mathbf{w}_1\cdot\mathbf{x}+t)\neq\mathrm{sign}(\mathbf{w}_2\cdot\mathbf{x}+t)]$$
$$\leq \frac{\theta(\mathbf{w}_1,\mathbf{w}_2)}{\pi}e^{-t^2/2}.$$

Observe now that since, by the assumptions of Lemma 3.3, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq y] > C\epsilon$ and $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(-\mathbf{w}\cdot\mathbf{x}+t)\neq y] > C\epsilon$, we obtain that θ can neither be very close to 0 or to π. In particular, we have that

$$\frac{1}{\pi}\theta(\mathbf{w},\mathbf{w}^*)e^{-t^2/2} \geq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq f(\mathbf{x})]$$
$$\geq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq y] - \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})\neq y]$$
$$\geq (C-1)\epsilon.$$

Similarly, we obtain that $\frac{1}{\pi}(\pi-\theta)e^{-t^2/2}\geq(C-1)\epsilon$, which implies that $\sin\theta e^{-t^2/2}/\epsilon$ is greater than the sufficiently large absolute constant $C-1$. It now suffices to show that it is larger than the "noise" term: $9\epsilon\sqrt{\log(q/\epsilon+1)}$. We can now replace the probability q by its upper bound and obtain

$$9\epsilon\sqrt{\log(q/\epsilon+1)} \leq 9\epsilon\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+3}$$
$$+ 9\epsilon\sqrt{(t\sin\theta)^2/2+|t|\sin\theta+3}.$$

We will first prove that the term $(2/3)\sin\theta e^{-t^2/2}/3$ is greater than $9\epsilon\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+3}$. Observe that since $\sin\theta e^{-t^2/2}/\epsilon\geq(C-1)>1$, it suffices to show that $\sin\theta e^{-t^2/2}/\epsilon$ is larger than $c\epsilon\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+1}$, for some absolute constant $c>0$. Using the inequality $t\geq r\sqrt{\log(t)+1}$, for $t\geq r^2\geq 1$, we obtain that when $\sin\theta e^{-t^2/2}/\epsilon\geq c^2$ the claim is true. Therefore, it suffices to make the constant C sufficiently large so that $C-1>c^2$. For the proof that $(2/3)\sin\theta e^{-t^2/2}/3 > 9\epsilon\sqrt{(t\sin\theta)^2/2+|t|\sin\theta+3}$ we refer to the Appendix.

□

Algorithm 1 Online Projected Gradient Descent (OPGD)

Input: Loss function ℓ, Step size λ, Current vector $\mathbf{w}$
$\mathbf{u} \leftarrow \mathbf{w} - \lambda \nabla \ell(\mathbf{w})$.
$\mathbf{w}' \leftarrow \mathrm{proj}_{\mathcal{B}}(\mathbf{u})$.
return $\mathbf{w}'$.

Algorithm 2 Online Projected Gradient Descent Learner

Input: $\epsilon \in (0,1)$, N examples $(\mathbf{x}^{(i)}, y^{(i)})$ from an ϵ-corrupted $\mathcal{D}$.

Denote $\widehat{\mathcal{D}}$ the N-sample empirical distribution of $\mathcal{D}$.
$t \leftarrow \Phi^{-1}\left(\frac{1-\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[y]}{2}\right)$.
$\mathbf{w}^{(0)} \leftarrow \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{x}y]$.
$s \leftarrow \Theta(1/\log^{3/2}(1/\epsilon))$.
$\lambda \leftarrow e^{t^2/2}s$ and $T \leftarrow \Theta(\log^4(1/\epsilon))$.
for $k = 0$ **to** T **do**
  $\phi_k \leftarrow (\pi/2)(1 - s^2/64)^k$.
  $a_k \leftarrow \sin(\phi_k)$.
  $b_k \leftarrow -t\cos(\phi_k)$.
  $\widehat{\mathcal{L}}_k(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[-r_{a_k,b_k}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right)y\right]$.

  Online Projected Gradient Descent Step
  $\mathbf{w}^{(k+1)} \leftarrow \mathrm{OPGD}\left(\widehat{\mathcal{L}}_k, \mathbf{w}^{(k)}, \lambda\right)$
end for
return $(\mathbf{w}^{(T)}, t)$.

4. Learning LTFs via Online Gradient Descent

In this section, we prove our main algorithmic result: we show that we can solve the non-convex feasibility problem of Section 3 using Online Gradient Descent; see Algorithm 2. The sequence of non-convex objectives that we use has the form $\mathcal{L}_{a,b}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[-r_{a,b}(\mathbf{w}\cdot\mathbf{x})y]$, where we define the ramp function $r_{a,b}$ with center b and length a as follows:

$$r_{a,b}(t) = \begin{cases} 0 & \text{if } t \le b-a \\ t-b+a & \text{if } b-a < t < b+a \\ 2a & \text{otherwise} \end{cases}$$

Similar "ramp" or sigmoidal loss functions have been previously used in (Diakonikolas et al., 2020b;c) in order to learn homogeneous halfspaces with label noise. In particular, in (Diakonikolas et al., 2020b) it was shown that finding a stationary points of the "centered" ramp activation, i.e., the loss $\mathcal{L}_{a,0}(\mathbf{w})$ for $a = \Theta(\epsilon)$, suffices to obtain a halfspace with error $O(\epsilon)$. While such fixed objectives cannot handle general halfspaces, we essentially show that following the gradients of a sequence of "ramp" objectives with different thresholds and lengths converges to an approximately optimal halfspace. At a high-level, the adversary of the online optimization process picks a loss function whose gradient violates some constraint of the non-convex problem (1) and the learning algorithm uses this gradient to improve its guess. We now formally state our result.

Theorem 4.1 (Online Gradient Descent Learner)**.** *Fix $\epsilon, \delta \in (0, 1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\widehat{\mathcal{D}}$ the empirical distribution formed with $N = \tilde{O}(\frac{d\log(1/\delta)}{\epsilon^2})$ samples from $\mathcal{D}$. Then the Online Gradient Descent Algorithm 2, after $T = O(\log^4(1/\epsilon))$ iterations, returns a vector $\mathbf{w}^{(T)}$ and a threshold t such that, with probability at least $1-\delta$, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(T)}\cdot\mathbf{x} + t) \neq y] \le C\epsilon$, where $C > 0$ is some universal constant.*

We first show that the gradient of $\mathcal{L}_k(\mathbf{w})$ is equal to $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[-y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \in B_{a,b}\}]$, i.e., it coincides (modulo a sign change) with the vector of the left hand side of the constraint of the non-convex problem (1).

Claim 4.2. *For any unit vector $\mathbf{w} \in \mathbb{R}^d$, the Online Projected Gradient update rule of Algorithm 1 with loss $\mathcal{L}_{a,b}(\mathbf{w})$ and stepsize λ corresponds to the update*

$$\mathbf{w}' = \frac{\mathbf{w} - \lambda\nabla_{\mathbf{w}}\mathcal{L}_{a,b}(\mathbf{w})}{\|\mathbf{w} - \lambda\nabla_{\mathbf{w}}\mathcal{L}_{a,b}(\mathbf{w})\|_2} .$$

Moreover, the gradient is equal to $\nabla_{\mathbf{w}}\mathcal{L}_{a,b}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[-y\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x} - b| \le a\}\nabla_{\mathbf{w}}\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right]$.

Proof. The gradient of $\mathcal{L}_{a,b}(\mathbf{w})$ is equal to

$$\begin{aligned}\nabla_{\mathbf{w}}\mathcal{L}_{a,b}(\mathbf{w}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[-r'_{a,b}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right)y\nabla_{\mathbf{w}}\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right] \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[-y\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x} - b| \le a\}\nabla_{\mathbf{w}}\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right] ,\end{aligned}$$

where in the first equality we used the chain rule. Observe that $\nabla_{\mathbf{w}}\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2} = \frac{1}{\|\mathbf{w}\|_2}(\mathbf{x} - \frac{(\mathbf{w}\cdot\mathbf{x})\mathbf{w}}{\|\mathbf{w}\|_2^2}) = \frac{(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})}{\|\mathbf{w}\|_2} = (\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})$, where we used that $\mathbf{w}$ is a unit norm vector. Hence $\nabla_{\mathbf{w}}\mathcal{L}_{a,b}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[-y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x} - b| \le a\}]$. Moreover, from the above we get that $\mathbf{w}\cdot\nabla_{\mathbf{w}}\mathcal{L}_{a,b}(\mathbf{w}) = 0$, therefore from the Pythagorean theorem we get that

$$\|\mathbf{w} - \lambda\nabla_{\mathbf{w}}\mathcal{L}_{a,b}(\mathbf{w})\|_2^2 = \|\mathbf{w}\|_2^2 + \lambda^2\|\nabla_{\mathbf{w}}\mathcal{L}_{a,b}(\mathbf{w})\|_2^2 \ge 1 ,$$

where we used that $\|\mathbf{w}\|_2 = 1$. Therefore, after each step, the Online Projected Gradient update rule will normalize the new vector so that it lies on the d-dimensional unit ball, i.e., the projection step of Algorithm 1 happens always. □

In Lemma 3.3 (see also Remark 3.4) we essentially showed that there exist parameters a, b such that the gradient $\nabla\mathcal{L}(\mathbf{w})$ "points to the direction" of $\mathbf{w}^*$. In particular, we showed that we can pick the band with $a = \sin\theta(\mathbf{w}, \mathbf{w}^*)$ and $b =$

$-t\cos(\theta(\mathbf{w},\mathbf{w}^*))$. Unfortunately, we do not know either the value of the optimal threshold t or the angle between the guess and the optimal weight vector $\mathbf{w}^*$: $\theta(\mathbf{w},\mathbf{w}^*)$. It is not hard to obtain an estimate t' of the value of the threshold t from the noisy samples: we can show (see Claim 4.7) that with $O(1/\epsilon^2)$ samples we can obtain a good estimate t'. In fact, we show that we can assume that t' is the optimal threshold and only introduce $O(\epsilon)$ additional noise in the distribution $\mathcal{D}$. Therefore, to keep the presentation clean, in what follows we will assume that we know the value of the optimal threshold t. One could hope that we can also estimate the angle between $\mathbf{w},\mathbf{w}^*$ from samples and assume that it is also known. It is unclear whether we can estimate the angle accurately enough: in fact, we will show that we do not need to do so. In general, we will need a "robust" version of Lemma 3.3 showing how close must be the threshold, and size values a,b to the "true" used in Lemma 3.3 in order for the gradient $\nabla\mathcal{L}_k(\mathbf{w})$ to point to the right direction. We prove the following lemma showing that it suffices to use the band $B_{a,b}$ with $a=\sin(\phi)$ and $b=-t\cos(\phi)$ assuming that $|\phi-\theta(\mathbf{w},\mathbf{w}^*)|\leq\phi/\log(1/\epsilon)$. The fact that we only have an (inverse) logarithmic tolerance (as opposed to requiring the difference to be $\mathrm{poly}(\epsilon)$) is crucial for obtaining the fast, i.e., in $\mathrm{poly}(\log(1/\epsilon))$ rounds, convergence of our Online Gradient Descent algorithm. In what follows, given some $\phi\in[0,\pi/2]$, we define the band

$$B_\phi=\{z\in\mathbb{R}:|z+t\cos(\phi)|<\sin(\phi)\}\,,$$

and the corresponding (negative) gradient of the loss function $\mathcal{L}_{\sin(\phi),-t\cos(\phi)}(\mathbf{w})$ as

$$\mathbf{g}(\mathbf{w},\phi)=\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_\phi\}]\,.$$

Lemma 4.3 ("Robust" Localized Gradient Update)**.** *Fix $\epsilon\in(0,1/2)$ and a sufficiently large constant $C>1$. Let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d\times\{\pm1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\mathbf{w}^*$ the weight vector of an optimal halfspace, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t)\neq y]\leq\epsilon$. Fix $\phi\in[0,\pi/2]$ and a unit vector $\mathbf{w}\in\mathbb{R}^d$ and assume that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq y]\geq C\epsilon$. Then, if $|\phi-\theta(\mathbf{w},\mathbf{w}^*)|\leq\phi/(C\log(1/\epsilon))$, it holds that*

$$\mathbf{g}(\mathbf{w},\phi)\cdot\mathbf{w}^*\geq(\sin\theta(\mathbf{w},\mathbf{w}^*))^2e^{-t^2/2}\,.$$

We now have shown that there exists a choice for the parameter ϕ of the ramp objective so that its gradient $g(\mathbf{w},\phi)$ points to the right direction. The following claim makes this fact precise and proves that when a vector $\mathbf{u}$ correlates positively with some target vector $\mathbf{v}^*$ and at the same time lies in the orthogonal complement of some current guess $\mathbf{u}$, then there exists a step size such that the normalized gradient step will improve the correlation with the target vector $\mathbf{v}^*$ (or equivalently decrease the angle with $\mathbf{v}^*$). Formally, we use the following claim (a variant of which was proved in (Diakonikolas et al., 2020a)).

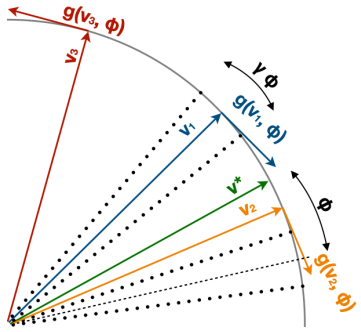

Figure 1. The "angle-contractive" map of Lemma 4.5. The target vector is $\mathbf{v}^*$ (green vector). Observe that for some fixed angle ϕ, as long as $|\theta(\mathbf{v},\mathbf{w}^*)-\phi|\leq\gamma\phi$. it holds that $\mathbf{g}(\mathbf{v},\phi)$ points to the right direction. This happens in the two dotted cones around $\mathbf{w}^*$. For example, for the blue vector $\mathbf{v}_1$ the gradient field $g(\mathbf{v}_1,\phi)$ points towards $\mathbf{v}^*$. Vectors outside of this region have noisy gradients that may point to any arbitrary direction. Notice that this may also happen when a vector is very close to $\mathbf{v}^*$. For example, even though the orange vector $\mathbf{v}_2$ is closer to $\mathbf{v}^*$ than $\mathbf{v}_1$, its gradient points in the wrong direction.

Claim 4.4 (Correlation Improvement (Diakonikolas et al., 2020a))**.** *For unit vectors $\mathbf{v}^*,\mathbf{v}\in\mathbb{R}^d$, let $\mathbf{u}\in\mathbb{R}^d$ such that $\mathbf{u}\cdot\mathbf{v}^*\geq c$, $\mathbf{u}\cdot\mathbf{v}=0$, and $\|\mathbf{u}\|_2\leq1$, with $c>0$. Then, for $\mathbf{v}'=\frac{\mathbf{v}+\lambda\mathbf{u}}{\|\mathbf{v}+\lambda\mathbf{u}\|_2}$, with $\lambda\leq c/2$, we have that $\mathbf{v}'\cdot\mathbf{v}^*\geq\mathbf{v}\cdot\mathbf{v}^*+\lambda c/8$.*

Given a current guess $\mathbf{w}$ we still have the issue of knowing the angle between $\mathbf{w}$ and $\mathbf{w}^*$. In the next lemma we show that we do not need to know the value of the angle but only an upper bound, i.e., we know that $\theta(\mathbf{w},\mathbf{w}^*)\in[0,\phi]$. We show that, with an appropriate step size, the online gradient descent update of Algorithm 1 is a contraction map in the sense that, even though an update may make the angle between the guess $\mathbf{w}$ and $\mathbf{w}^*$ worse, we can show that the new angle belongs in a smaller interval $[0,\phi']$ that is significantly smaller than our initial interval $[0,\phi]$.

Lemma 4.5 (Angle Contractive Map)**.** *Fix a unit vector $\mathbf{v}^*\in\mathbb{R}^d$, $\beta,\gamma,\kappa\in(0,1)$ and a vector field $\mathbf{g}:\mathbb{R}^d\times\mathbb{R}\mapsto\mathbb{R}^d$ such that for any vector $\mathbf{u}$ and $\phi\in[0,\pi/2]$, it holds that $\|\mathbf{g}(\mathbf{u},\phi)\|_2\leq\kappa$, $\mathbf{g}(\mathbf{u},\phi)\cdot\mathbf{u}=0$, and, if $|\phi-\theta(\mathbf{u},\mathbf{v}^*)|\leq\gamma\sin\phi$, and $\sin(\theta(\mathbf{u},\mathbf{v}^*))\geq\beta>0$ then $\mathbf{g}(\mathbf{u},\phi)\cdot\mathbf{v}^*\geq\rho\kappa\sin(\theta(\mathbf{u},\mathbf{v}^*))>0$. Fix $\phi\in[0,\pi/2]$ with $\sin(\phi)\geq\beta>0$. Set $\mathbf{v}$ to be any unit vector in $\mathbb{R}^d$ with $\theta(\mathbf{v},\mathbf{v}^*)\leq\phi$ and consider the normalized gradient update rule*

$$\mathbf{v}'=\frac{\mathbf{v}+\lambda\mathbf{g}(\mathbf{v},\phi)}{\|\mathbf{v}+\lambda\mathbf{g}(\mathbf{v},\phi)\|_2}\,,$$

with $\lambda=\rho\gamma\sin\phi/(4\kappa)$. Then, it holds that $\theta(\mathbf{v}',\mathbf{v}^*)\leq\phi'$, where, $\phi'=\max(\phi(1-\rho^2\gamma^2/64),\beta)$.

Remark 4.6. It is useful to provide some context and connect the parameters of Lemma 4.5 with the parameters of

Lemma 3.5. Assuming that $\mathbf{w}$ and $\mathbf{w}^*$ have angle θ, we can show that $\|\mathbf{g}(\mathbf{w},\phi)\| = O(\sin\theta e^{-t^2/2}\sqrt{\log(1/\epsilon)}) = \kappa$. Moreover, Lemma 3.5 shows that $\mathbf{g}(\mathbf{w},\phi)\cdot\mathbf{w}^* = \Omega(\kappa\sin\theta/\sqrt{\log(1/\epsilon)}) = \rho\kappa\sin\theta$ for some $\rho = \Omega(1/\sqrt{\log(1/\epsilon)})$. Moreover, we know that this holds as long as $|\phi-\theta| \le \sin\phi/(C\log(1/\epsilon))$. Therefore, $\gamma = \Omega(1/\log(1/\epsilon))$. Finally, β is the target angle which can be as small as $\Omega(\epsilon)$. Importantly, the decrease in the upper bound that we obtain in Lemma 4.5 is multiplicative: combining, these estimates we obtain that after one gradient update the angle upper bound $\phi' \le \phi(1-\Omega(1/\log^3(1/\epsilon)))$. Thus, in order to make sure that the angle lies in the interval $[0,\beta]$ we need at most $O(\log^4(1/\epsilon))$ iterations.

Proof. We shall distinguish two cases. In the first case we assume that $|\phi-\theta(\mathbf{v},\mathbf{v}^*)| \le \gamma\sin\phi$ (this corresponds to the vector $\mathbf{v}_1$ (blue) in Figure 1), and that means that it holds $\mathbf{g}(\mathbf{v},\phi)\cdot\mathbf{v}^* \ge \rho\sin(\theta(\mathbf{v},\mathbf{v}^*))/\kappa$, i.e., its gradient points in the right direction. From Claim 4.4, we have that $\mathbf{v}'\cdot\mathbf{v}^* \ge \mathbf{v}\cdot\mathbf{v}^* + \lambda\rho\sin(\theta(\mathbf{v},\mathbf{v}^*))/(8\kappa) \ge \mathbf{v}\cdot\mathbf{v}^* + \rho^2\gamma^2\phi^2/64$, where we used that $\lambda = \rho\gamma\sin\phi/(4\kappa) \ge \rho\gamma\phi/(8\kappa)$, because $\sin x \ge x/2$ for $\pi/2 \ge x \ge 0$. Hence, we have that $\cos(\theta(\mathbf{v}',\mathbf{v}^*)) \ge \cos(\theta(\mathbf{v},\mathbf{v}^*)) + \rho^2\gamma^2\phi^2/64$ and note that because $\cos(t)$ is decreasing in $[0,\pi]$, it holds that $\theta(\mathbf{v},\mathbf{v}^*) \ge \theta(\mathbf{v}',\mathbf{v}^*)$. Moreover, using the trigonometric identity $\cos x - \cos y = 2\sin((x+y)/2)\sin((y-x)/2)$ and that $\sin x \le x$ for $x > 0$, we get that

$$\theta(\mathbf{v},\mathbf{v}^*)^2 - \theta(\mathbf{v}',\mathbf{v}^*)^2 \ge \rho^2\gamma^2\phi^2/32\,,$$

and using that $\theta(\mathbf{v},\mathbf{v}^*) \le \phi$, we get that

$$\phi^2(1-\rho^2\gamma^2/32) \ge \theta(\mathbf{v}',\mathbf{v}^*)^2\,,$$

which completes the proof for this case, since we have shown that $\phi' = \phi(1-\rho^2\gamma^2/64) \ge \phi(1-\rho^2\gamma^2/32)^{1/2} \ge \theta(\mathbf{v}',\mathbf{v}^*)$.

We now assume that the true current angle $\theta(\mathbf{v},\mathbf{v}^*)$ is far from our current upper bound ϕ (this corresponds to vector $\mathbf{v}_2$ (orange) in Figure 1), i.e., $|\phi-\theta(\mathbf{v},\mathbf{v}^*)| \ge \gamma\sin\theta$. In this case, we will potentially do a step in the wrong direction, i.e., the angle between $\mathbf{v}'$ and $\mathbf{v}^*$ will become worse than before but still not worse than our new upper bound for the angle ϕ'. By Cauchy-Schwarz, and the fact that $\|\mathbf{g}(\mathbf{v},\phi)\|_2 \le \kappa$, we obtain $\|\mathbf{v}'-\mathbf{v}\|_2 \le 2\lambda\kappa$. Equivalently, we have that $|\cos(\theta(\mathbf{v}',\mathbf{v}^*)) - \cos(\theta(\mathbf{v},\mathbf{v}^*))| \le 2\lambda\kappa$. Using the fact that $t \mapsto \cos(t)$ is 1-Lipschitz, we obtain that $\theta(\mathbf{v}',\mathbf{v}^*) \le \theta(\mathbf{v},\mathbf{v}^*) + 2\kappa\lambda$. Since the initial angle was far from the upper bound ϕ, $\phi - \theta(\mathbf{v},\mathbf{v}^*) \ge \gamma\sin\phi$, we have:

$$\begin{aligned}\phi' - \theta(\mathbf{v}',\mathbf{v}^*) &\ge (\phi'-\phi) + (\phi-\theta(\mathbf{v},\mathbf{v}^*)) - 2\lambda\kappa \\ &\ge \gamma\sin\theta > 0\,.\end{aligned}$$

Therefore, we conclude that, in both cases, the new angle $\theta(\mathbf{v}',\mathbf{v}^*)$ belongs in the interval $[0,\phi']$. □

4.1. Proof Sketch of Theorem 4.1

We first show that the empirical threshold estimate obtained in the first step of Algorithm 2 is sufficiently accurate.

Claim 4.7 (Threshold estimation). *Fix $\epsilon,\delta \in (0,1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d\times\{\pm1\}$ with standard normal $\mathbf{x}$-marginal and let $\mathbf{w}^*, t$ be a unit vector and a threshold such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t) \neq y] \le \epsilon$. Denote $\widehat{\mathcal{D}}$ the empirical distribution with $N = O(\log(1/\delta)/\epsilon^2)$ samples and let $t' = \Phi^{-1}\left((1-\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[y])/2\right)$ be the empirical estimate of t used in Algorithm 2. It holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t') \neq y] \le O(\epsilon)$ with probability at least $1-\delta$.*

By Claim 4.7, using t' as the optimal threshold only introduces $O(\epsilon)$ additional noise in the distribution $\mathcal{D}$. Therefore, to keep the presentation clean, in what follows we will assume that we know the value of the optimal threshold t.

Observe that in Claim 4.4 and Lemma 4.3 we require a non-trivial initialization, namely that $\theta(\mathbf{w}^{(0)},\mathbf{w}^*) \in [0,\pi/2]$. We remark that choosing a random vector on the unit sphere will satisfy this assumption with probability $1/2$. However, we can do much better by choosing our initial vector to be $\mathbf{w}^{(0)} = \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{x}y]$. For simplicity, we ignore the sampling error, i.e., we consider the vector $\mathbf{w}^{(0)} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x}y]$ and we show that $\mathbf{w}^*\cdot\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x}y]$ is larger than $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$. Therefore, with sufficiently many samples (in particular $\widetilde{O}(d/\epsilon^2)$ suffice) we have that the empirical estimate of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x}y]$ will also have positive correlation with $\mathbf{w}^*$. Next, ignoring again sampling errors and assuming access to the population gradients of $\mathcal{L}_k(\mathbf{w})$, using Lemma 4.3 and Lemma 4.5, we have seen that with roughly $O(\log^4(1/\epsilon))$ updates we obtain that $\theta(\mathbf{w}^{(T)},\mathbf{w}^*) \le C\epsilon e^{-t^2/2}$; see also Remark 4.6. From Lemma 3.7 we obtain that $\mathbf{Pr}[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t) \neq \mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)] = O(\epsilon)$ and by a triangle inequality we conclude that also $\mathbf{Pr}[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t) \neq y] = O(\epsilon)$.

The analysis of the sample complexity of Algorithm 2 relies on standard concentration and union bound arguments and we defer it to the appendix. Finally, given N samples, the runtime of a single iteration of Algorithm 2 is $O(N)$ since we simply have to iterate over all samples and only keep those that fall in the band B_{a_k,b_k}. Since we are doing $T = O(\log^4(1/\epsilon))$ iterations the total runtime of Algorithm 2 is sample near-linear. We refer to Appendix C for more details.

Acknowledgements

Ilias Diakonikolas was supported by NSF Award CCF-1652862 (CAREER), a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant. Chris-

tos Tzamos and Vasilis Kontonis were supported by the NSF Award CCF-2144298 (CAREER). Nikos Zarifis was supported in part by NSF Award CCF-1652862 (CAREER) and a DARPA Learning with Less Labels (LwLL) grant.

References

Awasthi, P., Balcan, M. F., Haghtalab, N., and Urner, R. Efficient learning of linear separators under bounded noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pp. 167–190, 2015.

Awasthi, P., Balcan, M. F., Haghtalab, N., and Zhang, H. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pp. 152–192, 2016.

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

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005.

Daniely, A. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual Symposium on Theory of Computing, STOC 2016*, pp. 105–117, 2016.

Diakonikolas, I., Kane, D. M., and Stewart, A. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pp. 1061–1073, 2018.

Diakonikolas, I., Kane, D. M., Kontonis, V., Tzamos, C., and Zarifis, N. A polynomial time algorithm for learning halfspaces with tsybakov noise. *arXiv*, 2020a.

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

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

Frei, S., Cao, Y., and Gu, Q. Agnostic learning of a single neuron with gradient descent. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3a37abdeefe1dab1b30f7c5c7e581b93-Abstract.html.

Freund, Y. and Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Goldmann, M., Håstad, J., and Razborov, A. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. A short note on concentration inequalities for random vectors with subgaussian norm, 2019.

Kearns, M., Schapire, R., and Sellie, L. Toward Efficient Agnostic Learning. *Machine Learning*, 17(2/3):115–141, 1994.

Klivans, A. R. and Kothari, P. Embedding hard learning problems into gaussian space. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2014*, pp. 793–809, 2014.

Kouba, O. Inequalities related to the error function, 2006.

Maass, W. and Turan, G. How fast can a threshold gate learn? In Hanson, S., Drastal, G., and Rivest, R. (eds.), *Computational Learning Theory and Natural Learning Systems*, pp. 381–414. MIT Press, 1994.

Minsky, M. and Papert, S. *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, MA, 1968.

Naaman, M. On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. *Statistics & Probability Letters*, 173:109088, 2021. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2021.109088. URL https://www.sciencedirect.com/science/article/pii/S016771522100050X.

Novikoff, A. On convergence proofs on perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, volume XII, pp. 615–622, 1962.

O'Donnell, R. *Analysis of Boolean Functions*. Cambridge University Press, 2014. ISBN 978-1-10-703832-5.

Rosenblatt, F. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.

Shawe-Taylor, J. and Cristianini, N. *An introduction to support vector machines*. Cambridge University Press, 2000.

Shen, J. On the power of localized perceptron for label-optimal learning of halfspaces with adversarial noise, 2021.

Vapnik, V. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

Yan, S. and Zhang, C. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 1056–1066, 2017.

Yao, A. On ACC and threshold circuits. In *Proceedings of the Thirty-First Annual Symposium on Foundations of Computer Science*, pp. 619–627, 1990.

Zhang, C., Shen, J., and Awasthi, P. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

A. Preliminaries

We use small boldface characters for vectors. For $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, $\mathbf{x}_i$ denotes the i-th coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 \coloneqq (\sum_{i=1}^d \mathbf{x}_i^2)^{1/2}$ denotes the ℓ_2-norm of $\mathbf{x}$. We will use $\mathbf{x} \cdot \mathbf{y}$ for the inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\theta(\mathbf{x}, \mathbf{y})$ for the angle between $\mathbf{x}, \mathbf{y}$. For simplicity of notation, we may use θ instead of $\theta(\mathbf{x}, \mathbf{y})$ when it is clear from the context. We will use $\mathbb{1}_A$ to denote the characteristic function of the set A, i.e., $\mathbb{1}_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $\mathbb{1}_A(\mathbf{x}) = 0$ if $\mathbf{x} \notin A$. Let $\mathbf{e}_i$ be the i-th standard basis vector in $\mathbb{R}^d$. For a vector $\mathbf{w} \in \mathbb{R}^d$, we use $\mathbf{w}^\perp$ to denote the subspace spanned by vectors orthogonal to $\mathbf{w}$, i.e., $\mathbf{w}^\perp = \{\mathbf{u} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{u} = 0\}$. For a subspace $U \subseteq \mathbb{R}^d$, we denote $(\mathrm{proj}_U \mathbf{x})$, the projection of $\mathbf{x}$ onto U.

We use $\mathbf{E}_{x \sim \mathcal{D}}[x]$ for the expectation of the random variable x according to the distribution $\mathcal{D}$ and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. For simplicity of notation, we may omit the distribution when it is clear from the context. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the d-dimensional Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. we also denote $\mathcal{N}(\mu, \sigma^2)$, the standard normal distribution with mean μ and variance σ^2. For $(\mathbf{x}, y)$ distributed according to $\mathcal{D}$, we denote $\mathcal{D}_{\mathbf{x}}$ to be the distribution of $\mathbf{x}$ and $\mathcal{D}_y$ to be the distribution of y. For a set B and a distribution $\mathcal{D}$, we denote $\mathcal{D}_B$ to be the distribution $\mathcal{D}$ conditional on B. Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal, i.e., $\Phi(t) = 1/\sqrt{2\pi} \int_{-\infty}^t \exp(-z^2/2) \mathrm{d}z$, moreover, we denote $\Phi^{-1}(\cdot)$ to be the inverse function of $\Phi(\cdot)$.

B. Structural Result: A Non-Convex Feasibility Problem

In this section we prove our structural result, namely that there exists a non-convex feasibility problem whose solutions are approximately optimal halfspaces. In general, it is not hard to construct non-convex feasibility programs whose solutions are near-optimal vectors: in particular, minimizing the zero-one loss is indeed such a non-convex problem which is known to be computationally challenging under adversarial label noise even when the underlying distribution is the standard normal. Our non-convex feasibility formulation is inherently different than the standard zero-one loss minimization and its idenfiability proof is the basis of our Online Gradient Descent algorithm.

Theorem B.1 (Non-Convex Feasibility Program)**.** *Fix $\epsilon \in (0, 1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\mathbf{w}^*$ the weight vector of an optimal halfspace, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t) \neq y] \leq \epsilon$. Denote $B_{a,b} = \{z \in \mathbb{R} : |z - b| \leq a\}$ and consider the following feasibility program*

$$\begin{aligned} &\textit{Find} && \mathbf{w} : \|\mathbf{w}\|_2 = 1 \\ &\textit{s. t.} && \left\| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [y(\mathrm{proj}_{\mathbf{w}^\perp} \mathbf{x}) \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B_{a,b}\}] \right\|_2 \leq 4\sqrt{e}\epsilon \sqrt{\log\left(\mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)} [z \in B_{a,b}]/\epsilon + 1 \right)} \quad \forall a \geq 0, b \in \mathbb{R} \end{aligned} \tag{3}$$

We have that: (i) the above program is feasible and (ii) any solution $\mathbf{w}$ satisfies $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] \leq C\epsilon$ or $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(-\mathbf{w} \cdot \mathbf{x} + t) \neq y] \leq C\epsilon$, where C is some universal constant.

Before we prove the theorem, we would like to highlight its connection with our algorithmic result. Our Online Gradient Descent algorithm essentially uses as gradients the vectors in the left-hand side of the constraint of the non-convex program Equation (3). In Appendix C we will show that, as long as the current halfspace h is not nearly optimal, we can find an appropriate band $B_{a,b}$ and use the vector $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp} \mathbf{x}) \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B_{a,b}\}]$ to improve its weight vector.

We split the proof of Theorem B.1 in two parts: in Lemma B.2, where we show that the non-convex program is feasible and Lemma B.5, where we show that any of its solutions is an approximately optimal halfspace. We will show that any unit vector $\mathbf{w} \in \mathbb{R}^d$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] \leq C\epsilon$, where $C > 0$ is some sufficiently large universal constant, satisfies all the constraints of (3). We prove the following lemma.

Lemma B.2 (Feasibility)**.** *Fix $\epsilon \in (0, 1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\mathbf{w}^*$ the weight vector of an optimal halfspace, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t) \neq y] \leq \epsilon$. Fix unit vector $\mathbf{w} \in \mathbb{R}^d$ with $\theta(\mathbf{w}, \mathbf{w}^*) = \theta$ and denote $\mathbf{g} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp} \mathbf{x}) \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B_{a,b}\}]$ with $B_{a,b} = \{z \in \mathbb{R} : |z - b| \leq a\}$. It holds*

$$\|\mathbf{g}\|_2 \leq O(\epsilon + \theta e^{-t^2/2}) \sqrt{\log\left(\mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)} [z \in B_{a,b}]/\epsilon \right)}.$$

Moreover, if $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] \leq \epsilon$, it holds that

$$\|\mathbf{g}\|_2 \leq 4\sqrt{e}\epsilon \sqrt{\log\left(\mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)} [z \in B_{a,b}]/\epsilon \right)}.$$

Remark B.3. In particular, Lemma B.2 shows that $\mathbf{w}^*$ is a solution of the non-convex feasibility system (3). We remark that by relaxing the constraint of the non-convex program to $\left\|\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\left\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\right\}\right\|_2 \leq 4\sqrt{e}C'\epsilon\sqrt{\log\left(\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z\in B_{a,b}]/\epsilon+1\right)}$ for some larger constant $C'>1$ we will obtain that any halfspace h with error $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq y]\leq C'\epsilon$ will also be a feasible solution.

Proof. Note that $\|\mathbf{g}\|_2=\sup_{\mathbf{v}\in\mathbb{R}^d}|\frac{\mathbf{v}}{\|\mathbf{v}\|_2}\cdot\mathbf{g}|$. Pick any unit vector $\mathbf{u}\in\mathbb{R}^d$ and denote $h(\mathbf{x})=\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)$. We have that

$$\begin{aligned}|\mathbf{u}\cdot\mathbf{g}| &\leq |\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-h(\mathbf{x}))\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}]|+|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x})\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}]|\\ &=|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-h(\mathbf{x}))\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}]|\leq 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b},y\neq h(\mathbf{x})\}]\,,\end{aligned}$$

where for the first inequality we used triangle inequality; in the equality we used the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x})\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}]=\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})]\,\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}]=0$ because $\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})$ is independent of $h(\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b}\}$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})]=0$ because $\mathcal{D}_\mathbf{x}$ is zero-mean; and in the last inequality we used that $|y-h(\mathbf{x})|$ is non-zero only when $y\neq h(\mathbf{x})$ and at most 2. We now have to bound from above the contribution of the term $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B_{a,b},y\neq h(\mathbf{x})\}]$. One could use the Cauchy-Schwarz inequality to bound this expectation by $(\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u}\cdot(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})|^2])^{1/2}(\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{w}\cdot\mathbf{x}\in B_{a,b},y\neq h(\mathbf{x})])^{1/2}$. However, this would only imply an upper bound of the order of $(\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{w}\cdot\mathbf{x}\in B_{a,b},y\neq h(\mathbf{x})])^{1/2}=O(\sqrt{\epsilon})$, where we used that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[y\neq h(\mathbf{x})]\leq\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[y\neq h(\mathbf{x})]+\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[y\neq f(\mathbf{x})]=O(\epsilon)$. Using the concentration of the Gaussian distribution and the fact that $\mathbf{u}\cdot\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x}$ is independent from $\mathbf{w}\cdot\mathbf{x}$ we are able to prove a much stronger decoupling inequality. We show the following lemma.

Lemma B.4 (Gaussian Decoupling Inequality)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$ with standard normal $\mathbf{x}$-marginal. Moreover, let $\mathbf{w},\mathbf{u}\in\mathbb{R}^d$ be two orthogonal unit vectors, define $B=\{z\in\mathbb{R}:z\in(t_1,t_2)\}$, for some $t_1,t_2\in\mathbb{R}$ and let $S(\mathbf{x},y)$ be an event over $\mathbb{R}^d\times\{\pm1\}$. It holds that*

$$\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{S(\mathbf{x},y),\mathbf{w}\cdot\mathbf{x}\in B\}]\leq 2\sqrt{e}\,\mathbf{Pr}[S(\mathbf{x},y)]\sqrt{\log\left(\frac{\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x}\in B]}{\mathbf{Pr}[S(\mathbf{x},y)]}+1\right)}\,.$$

Proof. We first observe that the inequality is trivially true when $\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x}\in B]=0$. To simplify notation we will not now condition on the event $\mathbf{w}\cdot\mathbf{x}\in B$. For any $\xi>0$, we have that

$$\begin{aligned}&\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{S(\mathbf{x},y)\}\mid\mathbf{w}\cdot\mathbf{x}\in B]\\ &=\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{S(\mathbf{x},y),|\mathbf{u}\cdot\mathbf{x}|\leq\xi\}\mid\mathbf{w}\cdot\mathbf{x}\in B]+\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{S(\mathbf{x},y),|\mathbf{u}\cdot\mathbf{x}|\geq\xi\}\mid\mathbf{w}\cdot\mathbf{x}\in B]\\ &\leq\xi\,\mathbf{Pr}[S(\mathbf{x},y)\mid\mathbf{x}\in B]+\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq\xi\}\mid\mathbf{w}\cdot\mathbf{x}\in B]\,.\end{aligned}$$

We now observe that since $\mathbf{w}$ and $\mathbf{u}$ are orthogonal the Gaussian random variables $\mathbf{w}\cdot\mathbf{x}$ and $\mathbf{u}\cdot\mathbf{x}$ are independent and therefore, we can compute the second expectation as follows:

$$\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq\xi\}\mid\mathbf{w}\cdot\mathbf{x}\in B]=\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq\xi\}]=\frac{1}{\sqrt{2\pi}}\int_\xi^{+\infty}ze^{-z^2/2}=\frac{e^{-\xi^2/2}}{\sqrt{2\pi}}\,.$$

Denote $p=\mathbf{Pr}[S(\mathbf{x},y)\mid\mathbf{w}\cdot\mathbf{x}\in B]$ and set $\xi=\sqrt{2\log(1/p)}$. We obtain that

$$\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{S(\mathbf{x},y)\}\mid\mathbf{w}\cdot\mathbf{x}\in B]\leq p\sqrt{2\log(1/p)}+\frac{p}{\sqrt{2\pi}}\,.$$

Using the elementary inequalities $\sqrt{a}+\sqrt{b}\leq\sqrt{2}\sqrt{a+b}$ which is true for all $a,b>0$ and $\log z+1\leq e\log(z+1)$, which is true for all $z>0$, we obtain $\sqrt{2\log(1/p)}+1/\sqrt{2\pi}\leq 2\sqrt{e}\sqrt{\log(1/p+1)}$ for every $p\in(0,1]$.

Using that $\mathbf{Pr}[S(\mathbf{x},y)\mid\mathbf{w}\cdot\mathbf{x}\in B]=\mathbf{Pr}[S(\mathbf{x},y)\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B\}]/\,\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x}\in B]$ we obtain

$$\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{S(\mathbf{x},y),\mathbf{w}\cdot\mathbf{x}\in B\}]\leq 2\sqrt{e}\,\mathbf{Pr}[S(\mathbf{x},y),\mathbf{w}\cdot\mathbf{x}\in B]\sqrt{\log\left(\frac{\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x}\in B]}{\mathbf{Pr}[S(\mathbf{x},y),\mathbf{w}\cdot\mathbf{x}\in B]}\right)}\,.$$

Moreover, it is not hard to see that for all $r > 0$ the function $p \mapsto p\sqrt{\log(r/p+1)}$ is increasing in $p > 0$ and therefore we can replace the probability $\mathbf{Pr}[S(\mathbf{x}, y), \mathbf{w} \cdot \mathbf{x} \in B]$ by its upper bound $\mathbf{Pr}[S(\mathbf{x}, y)]$ in the previous bound to obtain

$$\mathbf{E}[|\mathbf{u} \cdot \mathbf{x}| \mathbb{1}\{S(\mathbf{x}, y), \mathbf{w} \cdot \mathbf{x} \in B\}] \leq 2\sqrt{e}\, \mathbf{Pr}[S(\mathbf{x}, y)] \sqrt{\log\left(\frac{\mathbf{Pr}[\mathbf{w} \cdot \mathbf{x} \in B]}{\mathbf{Pr}[S(\mathbf{x}, y)]} + 1\right)}\,.$$

□

Using Lemma B.4, we get that

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u} \cdot (\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})| \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B_{a,b}, y \neq h(\mathbf{x})\}] &\leq 2\sqrt{e} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \sqrt{\log\left(\frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \in B_{a,b}]}{\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]}\right)} \\ &\leq O(\epsilon + \theta e^{-t^2/2}) \sqrt{\log\left(\frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \in B_{a,b}]}{\epsilon}\right)}\,,\end{aligned}$$

where we used that $\epsilon \leq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] \leq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y] + \mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})] = \epsilon + O(\theta e^{-t^2/2})$. To bound the disagreement between $h(\mathbf{x})$ and the optimal halfspace $f(\mathbf{x})$ we used Fact C.11. From the same computation as above, we conclude that when $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] \leq \epsilon$, it holds that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u} \cdot (\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})| \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B_{a,b}, y \neq h(\mathbf{x})\}] \leq 4\sqrt{e}\epsilon \sqrt{\log\left(\frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \in B_{a,b}]}{\epsilon}\right)}\,,$$

which proves the second part of the lemma. □

Next, we show that for any vector $\mathbf{w}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] \geq C\epsilon$, where $C > 0$ is some sufficiently large universal constant, there exists a set B which violates a constraint of the non-convex program (3). In particular we show that choosing $a = \sin(\theta(\mathbf{w}, \mathbf{w}^*))$ and $b = -t\cos(\theta(\mathbf{w}, \mathbf{w}^*))$ we get a violated constraint of problem (3). We show the following lemma.

Lemma B.5. *Fix $\epsilon \in (0, 1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\mathbf{w}^*$ the weight vector of an optimal halfspace, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t) \neq y] \leq \epsilon$. Fix unit vector $\mathbf{w} \in \mathbb{R}^d$ and assume that $\min(e^{-t^2/2}\sin(\theta(\mathbf{w}, \mathbf{w}^*)), e^{-t^2/2}/|t|) \geq C\epsilon$, where $C > 0$ is some sufficiently large universal constant. Denote $\mathbf{g} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B_{a,b}\}]$ and pick $a = \sin\theta(\mathbf{w}, \mathbf{w}^*)$ and $b = -t\cos\theta(\mathbf{w}, \mathbf{w}^*)$. Then, it holds that,*

$$\|\mathbf{g}\|_2 \geq (4\sqrt{e})\, \epsilon \sqrt{\log\left(\mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[B_{a,b}]/\epsilon\right)}\,.$$

Remark B.6. We remark that, in fact, our proof of Lemma B.5 establishes a stronger condition which will eventually allow us to design an efficient algorithm. We show that the vector $\mathbf{g} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B_{a,b}\}]$ correlates positively with the optimal vector $\mathbf{w}^*$, i.e., that $\mathbf{g} \cdot \mathbf{w}^* > 0$. At a high-level this means that $\mathbf{g}$ "points to the right direction" and we can use it in order to improve our current guess, see Lemma B.7 and Section C.

Proof. As we plant to bound from below the norm of $\mathbf{g}$ it suffices to consider any unit vector $\mathbf{u}$ and show that $\mathbf{g} \cdot \mathbf{u}$ is large. We will first prove a general lemma that given any band B, bounds from below the inner product of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})y\mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B\}]$ with the direction $\mathbf{w}^*$.

Lemma B.7. *Fix $\epsilon \in (0, 1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\mathbf{w}^*$ with $\|\mathbf{w}^*\|_2 = 1$, the weight vector of an optimal halfspace, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t) \neq y] \leq \epsilon$. Fix some unit vector $\mathbf{w} \in \mathbb{R}^d$ such that $\theta(\mathbf{w}, \mathbf{w}^*) = \theta \in (0, \pi)$ and let $B = \{z \in \mathbb{R} : t_1 < z < t_2\}$. Denote $\mathbf{g} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in B\}]$ and $\mathbf{v} = \frac{\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*}{\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2}$. It holds*

$$\mathbf{g} \cdot \mathbf{v} \geq \sqrt{\frac{2}{\pi}}\left(\sin\theta e^{-\frac{t^2}{2}} p - 9\epsilon\sqrt{\log(q/\epsilon + 1)}\right)\,,$$

where $p = \mathbf{Pr}_{z\sim\mathcal{N}(-t\cos\theta, (\sin\theta)^2)}[z \in B]$, and $q = \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \in B]$.

We are now ready to prove Lemma B.5. Recall, that in Lemma B.5 we pick the band $B_{a,b}$ with $a = \sin(\theta(\mathbf{w}, \mathbf{w}^*))$ and $b = -t\cos(\theta(\mathbf{w}, \mathbf{w}^*))$. We have that

$$p = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(-t\cos\theta,(\sin\theta)^2)}[z \in B] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,(\sin\theta)^2)}[|z| \le \sin\theta] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[|z| \le 1] \ge \frac{2}{3}\,.$$

Moreover, we have

$$\begin{aligned} q &= \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[z \in B] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[|z + t\cos\theta| \le \sin\theta] \le 2\sin\theta \max(e^{-(-t\cos\theta+\sin\theta)^2/2}, e^{-(-t\cos\theta-\sin\theta)^2/2}) \\ &\le 2\sin\theta e^{-(t\cos\theta)^2/2+|t|\sin\theta+1/2} \le 4\sin\theta e^{-t^2/2}\, e^{(t\sin\theta)^2/2+|t|\sin\theta}\,. \end{aligned}$$

Therefore, we have that $\log(q/\epsilon+1) \le \log(\sin\theta e^{-t^2/2}/\epsilon) + (t\sin\theta)^2/2 + |t|\sin\theta + 3$. Using Lemma B.7, and the inequality $\sqrt{a}+\sqrt{b} \le \sqrt{2}\sqrt{a+b}$, we have that

$$\begin{aligned} \frac{1}{\sin\theta}(\mathrm{proj}_{\mathbf{w}^\perp}\bar{\mathbf{g}})\cdot\mathbf{w}^* &\ge \sqrt{\frac{2}{\pi}}\left(\sin\theta e^{-\frac{t^2}{2}}p - 9\epsilon\sqrt{\log(q/\epsilon+1)}\right) \\ &\ge \frac{2\sqrt{2}}{3\sqrt{\pi}}\left(\sin\theta e^{-\frac{t^2}{2}} - 20\epsilon\big(\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+3} + \sqrt{(t\sin\theta)^2+|t|\sin\theta}\big)\right)\,, \end{aligned}$$

Observe now that since, by the assumptions of Lemma C.3, it holds that $\sin\theta e^{-t^2/2}/\epsilon$ is greater than some sufficiently large absolute constant $C > 1$, it suffices to show that it is larger than each of the "noise" terms: $\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+3}$, $\sqrt{(t\sin\theta)^2+|t|\sin\theta}$, separately. We will first prove that the term $\sin\theta e^{-t^2/2}/3$ is greater than $20\epsilon\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+3}$. Observe that since $\sin\theta e^{-t^2/2} \ge 1$ it suffices to show that it is larger than $20\sqrt{3}\epsilon\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+1}$. We will use the following elementary inequality.

Claim B.8. *Let $c \ge 1$. Then for all $t \ge c^2$ it holds that $t \ge c\sqrt{\log(t)+1}$.*

Proof. Follows immediately from the inequality $\log(t) \le t-1$. □

Using the above claim, we obtain that when $\sin\theta e^{-t^2/2}/\epsilon \ge (20\cdot 3\sqrt{3})^2$ it holds that $\sin\theta e^{-t^2/2}/3 \ge 20\sqrt{3}\epsilon\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+1}$. We next show that $\sin\theta e^{-t^2/2}/3 \ge 20\epsilon\sqrt{(t\sin\theta)^2+|t|\sin\theta}$. We distinguish two cases. First, we handle the case $|t|\sin\theta \le 1$. In this case, we have to show that $\sin\theta e^{-t^2/2}/3 \ge 20\sqrt{6}\epsilon$ which holds directly from assumptions. When $|t|\sin\theta \ge 1$ we have that $\sqrt{(t\sin\theta)^2+|t|\sin\theta} \le \sqrt{2}|t|\sin\theta$. Therefore, in order for $\sin\theta e^{-t^2/2}/3$ to be larger than $20\sqrt{2}\epsilon|t|\sin\theta$ we need $e^{-t^2/2}/|t| \ge 3\cdot 20\sqrt{2}\epsilon$, which holds from the assumption that $e^{-t^2/2}/(\epsilon|t|)$ is greater than an absolute constant.

□

B.1. The Proof of Lemma B.7

Without loss of generality, to simplify notation, we assume that $\mathbf{w} = \mathbf{e}_2$ and $\mathbf{w}^* = -\sin\theta\mathbf{e}_1 + \cos\theta\mathbf{e}_2$. We have that $\mathbf{v} = -\mathbf{e}_1$. Let $S(\mathbf{x}, y)$ be the event that a point $(\mathbf{x}, y)$ is corrupted and note that $\mathbf{E}[\mathbb{1}\{S(\mathbf{x}, y)\}] = \epsilon$. Therefore, we have

$$\begin{aligned} \mathbf{g}\cdot\mathbf{v} &= \mathbf{E}[-\mathbf{x}_1 y\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B\}] \\ &= \mathbf{E}[-\mathbf{x}_1\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t)\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B\}] - 2\,\mathbf{E}[-|\mathbf{x}_1|\mathbb{1}\{S(\mathbf{x},y)\}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B\}] \\ &= \underbrace{\mathbf{E}[|\mathbf{x}_1|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B, \mathbf{x}_1(\mathbf{w}^*\cdot\mathbf{x}+t)<0\}] - \mathbf{E}[|\mathbf{x}_1|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B, \mathbf{x}_1(\mathbf{w}^*\cdot\mathbf{x}+t)\ge 0\}]}_{I_1} \\ &\quad - 2\underbrace{\mathbf{E}[|\mathbf{x}_1|\mathbb{1}\{S(\mathbf{x},y)\}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B\}]}_{I_2}\,. \end{aligned}$$

Using the Lemma B.4, we get that

$$I_2 \le 2\sqrt{e}\epsilon\sqrt{\log(\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x}\in B]/\epsilon+1)}\,, \tag{4}$$

where we used that $\mathbf{Pr}[S] = \epsilon$. Next, we prove the following claim for the term I_1.

Claim B.9. *It holds that*

$$I_1 = \frac{\sin\theta}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(-t\cos\theta,\sin\theta^2)}[t_1 < x < t_2] \,.$$

Proof. First we show that

$$I_1 = 2\,\mathbf{E}\left[|\mathbf{x}_1|\left\{\mathbf{w}\cdot\mathbf{x}\in B, \mathbf{x}_1 \geq \frac{|\mathbf{x}_2\cos\theta + t|}{\sin\theta}\right\}\right] \,.$$

Fix $t_1 \leq \mathbf{x}_2 \leq t_2$ and we have two cases. The first case is when $\mathbf{x}_2\cos\theta + t \geq 0$, then $\mathbf{x}_1(\mathbf{w}^*\cdot\mathbf{x}+t)\leq 0$ is equivalent to $\mathbf{x}_1 \in (-\infty, 0)\cup((\mathbf{x}_2\cos\theta+t)/\sin\theta, \infty)$. The second case is when $\mathbf{x}_2\cos\theta + t < 0$, then $\mathbf{x}_1(\mathbf{w}^*\cdot\mathbf{x}+t)\leq 0$ is equivalent to $\mathbf{x}_1 \in (-\infty, -(\mathbf{x}_2\cos\theta+t)/\sin\theta)\cup(0,\infty)$. Note that from the symmetry of the Gaussian distribution, i.e., $\mathbf{x}_1$ has the same distribution with $-\mathbf{x}_1$, we have

$$\mathbf{E}[|\mathbf{x}_1|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B, \mathbf{x}_1(\mathbf{w}^*\cdot\mathbf{x}+t)<0\}] = \mathbf{E}[|\mathbf{x}_1|\{\mathbf{w}\cdot\mathbf{x}\in B, \mathbf{x}_1\in(-\infty,0)\cup(|\mathbf{x}_2\cos\theta+t|/\sin\theta,\infty)\}] \,,$$

and note that $\mathbf{E}[|\mathbf{x}_1|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B, \mathbf{x}_1(\mathbf{w}^*\cdot\mathbf{x}+t)\geq 0\}] = \mathbf{E}[|\mathbf{x}_1|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B\}] - \mathbf{E}[|\mathbf{x}_1|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in B, \mathbf{x}_1(\mathbf{w}^*\cdot\mathbf{x}+t)<0\}]$ and that $\mathbf{E}[|\mathbf{x}_1|\{\mathbf{w}\cdot\mathbf{x}\in B, \mathbf{x}_1\in(-\infty,0)\}] = (1/2)\,\mathbf{E}[|\mathbf{x}_1|\{\mathbf{w}\cdot\mathbf{x}\in B\}]$, therefore,

$$I_1 = 2\,\mathbf{E}\left[|\mathbf{x}_1|\left\{\mathbf{w}\cdot\mathbf{x}\in B, \mathbf{x}_1 \geq \frac{|\mathbf{x}_2\cos\theta + t|}{\sin\theta}\right\}\right] \,.$$

Next, we have that

$$I_1 = \frac{1}{2\pi}\int_{t_1}^{t_2}\int_{\frac{|t+\cos\theta\mathbf{x}_2|}{\sin\theta}}^{\infty} \mathbf{x}_1\exp(-(\mathbf{x}_1^2-\mathbf{x}_2^2)/2)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2 = \frac{1}{2\pi}\int_{t_1}^{t_2}\exp\left(-\left(\frac{t}{\sin\theta}+\frac{\mathbf{x}_2}{\tan\theta}\right)^2/2 - \mathbf{x}_2^2/2\right)\mathrm{d}\mathbf{x}_2 \,.$$

Next, we prove the following claim

Claim B.10. *For $a \geq 0$ and $b\in\mathbb{R}$, it holds that*

$$\int_{t_1}^{t_2}\exp\left(-ax^2+bx\right)\mathrm{d}x = \frac{\exp(b^2/(4a))\sqrt{\pi}}{\sqrt{a}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(b/(2a),1/(2a))}[t_1 < x < t_2] \,.$$

Proof. We have that

$$\begin{aligned}\int_{t_1}^{t_2}\exp\left(-ax^2+bx\right)\mathrm{d}x &= \int_{t_1}^{t_2}\exp\left(-a(x-b/(2a))^2 + b^2/(4a)\right)\mathrm{d}x \\ &= \exp(b^2/(4a))\int_{t_1}^{t_2}\exp\left(-\frac{1}{2}(2a)(x-\frac{b}{2a})^2\right)\mathrm{d}x \,,\end{aligned}$$

Therefore, we have that

$$\int_{t_1}^{t_2}\exp\left(-ax^2+bx\right)\mathrm{d}x = \frac{\exp(b^2/(4a))\sqrt{\pi}}{\sqrt{a}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(b/(2a),1/(2a))}[t_1 < x < t_2] \,.$$

□

Using the claim above, we have that

$$\begin{aligned}&\frac{1}{2\pi}\int_{t_1}^{t_2}\exp(-\left(\frac{t}{\sin\theta}+\frac{\mathbf{x}_2}{\tan\theta}\right)^2/2 - \mathbf{x}_2^2/2)\mathrm{d}\mathbf{x}_2 \\ &\qquad = \frac{\exp(-\frac{1}{2}(t/\sin\theta)^2)}{2\pi}\int_{t_1}^{t_2}\exp\left(-\frac{1}{2}(1+\frac{1}{\tan^2\theta})\mathbf{x}_2^2 - \mathbf{x}_2 t\frac{\cos\theta}{\sin^2\theta}\right)\mathrm{d}\mathbf{x}_2 \,,\end{aligned}$$

therefore for $a = \frac{1}{2}(1 + 1/\tan\theta^2) = \frac{1}{2}(\sin\theta)^{-2}$ and $b = -t\cos\theta/\sin^2\theta$, we have that

$$\frac{1}{2\pi}\int_{t_1}^{t_2} \exp\left(-\left(\frac{t}{\sin\theta} + \frac{\mathbf{x}_2}{\tan\theta}\right)^2/2 - \mathbf{x}_2^2/2\right) \mathrm{d}\mathbf{x}_2 = \frac{\sin\theta}{\sqrt{2\pi}} e^{-\frac{t^2}{2}\left((1/\sin\theta)^2 - \cos\theta^2/\sin\theta^2\right)} \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(-t\cos\theta, \sin\theta^2)}[t_1 < x < t_2]$$
$$= \frac{\sin\theta}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(-t\cos\theta, \sin\theta^2)}[t_1 < x < t_2] \,.$$

□

Using Claim B.9 and Equation (4), we get

$$\mathbf{E}[-\mathbf{x}_1 y \mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \in B\}] \geq 2\left(\frac{\sin\theta}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(-t\cos\theta, \sin\theta^2)}[z \in B] - 2\sqrt{e}\epsilon\sqrt{\log(\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x} \in B]/\epsilon + 1)}\right) \,.$$

Using that $2\sqrt{2e\pi} \leq 9$, we get that

$$\mathbf{E}[-\mathbf{x}_1 y \mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \in B\}] \geq \sqrt{\frac{2}{\pi}}\left(\sin\theta e^{-\frac{t^2}{2}} \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(-t\cos\theta, \sin\theta^2)}[z \in B] - 9\epsilon\sqrt{\log(\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x} \in B]/\epsilon + 1)}\right) \,,$$

which completes the proof.

C. Learning LTFs via Online Gradient Descent

In this section we prove our main algorithmic result: we show that we can solve the non-convex feasibility problem of Appendix B using Online Gradient Descent. We first formally state our result.

Theorem C.1 (Non-Adaptive Online Gradient Descent Learner)**.** *Fix $\epsilon, \delta \in (0, 1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\widehat{\mathcal{D}}$ the empirical distribution formed with $N = \tilde{O}(\frac{d\log(1/\delta)}{\epsilon^2})$ samples from $\mathcal{D}$. Then, the Online Gradient Descent Algorithm 2, after $T = O(\log^4(1/\epsilon))$ iterations, returns a vector $\mathbf{w}^{(T)}$ and a threshold t such that, with probability at least $1-\delta$, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(T)}\cdot\mathbf{x} + t)) \neq y] \leq C\epsilon$, where $C > 0$ is some universal constant.*

Recall that the sequence of non-convex objectives that we use has the form $\mathcal{L}_k(\mathbf{w}) = -\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[r_{a_k,b_k}(\mathbf{w}\cdot\mathbf{x}/\|\mathbf{w}\|_2)y]$, where we define the ramp function $r_{a,b}$ with center b and length a as follows:

$$r_{a,b}(t) = \begin{cases} 0 & \text{if } t \leq b - a \\ t - b + a & \text{if } b - a < t < b + a \\ 2a & \text{otherwise} \end{cases}$$

We first show that the gradient of $\mathcal{L}_k(\mathbf{w})$ is equal to $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[-y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\mathbf{w}\cdot\mathbf{x} \in B_{a,b}]$, i.e., it coincides with the vector of the left hand side of the constraints of the non-convex problem (1).

Claim C.2. *For any unit vector $\mathbf{w} \in \mathbb{R}^d$, the Online Projected Gradient update rule of Algorithm 1 with loss $\mathcal{L}_k(\mathbf{w})$ and stepsize λ corresponds to the update*

$$\mathbf{w}' = \frac{\mathbf{w} - \lambda\nabla_{\mathbf{w}}\mathcal{L}_k(\mathbf{w})}{\|\mathbf{w} - \lambda\nabla_{\mathbf{w}}\mathcal{L}_k(\mathbf{w})\|_2} \,.$$

Moreover, $\nabla_{\mathbf{w}}\mathcal{L}_k(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[-y\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x} - b_k| \leq a_k\}\nabla_{\mathbf{w}}\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right]$.

Proof. The gradient of $\mathcal{L}_k(\mathbf{w})$ is equal to

$$\nabla_{\mathbf{w}}\mathcal{L}_k(\mathbf{w}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[-\nabla_{\mathbf{w}} r_{a_k,b_k}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right)y\right] = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[-r'_{a_k,b_k}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right)y\nabla_{\mathbf{w}}\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right]$$
$$= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[-y\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x} - b_k| \leq a_k\}\nabla_{\mathbf{w}}\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2}\right] \,,$$

where in the first equality we used the chain rule. Observe that $\nabla_{\mathbf{w}} \frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{w}\|_2} = \frac{1}{\|\mathbf{w}\|_2}(\mathbf{x} - \frac{(\mathbf{w}\cdot\mathbf{x})\mathbf{w}}{\|\mathbf{w}\|_2^2}) = \frac{(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})}{\|\mathbf{w}\|_2} = (\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})$, where we used that $\mathbf{w}$ is a unit norm vector. Hence $\nabla_{\mathbf{w}}\mathcal{L}_k(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[-y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x} - b_k| \le a_k\}]$. Moreover, from the above we get that $\mathbf{w}\cdot\nabla_{\mathbf{w}}\mathcal{L}_k(\mathbf{w}) = 0$, therefore from Pythagorian theorem we get that

$$\|\mathbf{w} - \lambda\nabla_{\mathbf{w}}\mathcal{L}_k(\mathbf{w})\|_2^2 = \|\mathbf{w}\|_2^2 + \lambda^2\|\nabla_{\mathbf{w}}\mathcal{L}_k(\mathbf{w})\|_2^2 = 1 + \lambda^2\|\nabla_{\mathbf{w}}\mathcal{L}_k(\mathbf{w})\|_2^2 \ge 1\,,$$

where we used that $\|\mathbf{w}\|_2 = 1$. Therefore, after each step, the Online Projected Gradient update rule will project the new vector inside the d-dimensional unit ball, hence the result follows. □

In Lemma B.5 (see also Remark B.6) we essentially showed that there exist parameters a, b such that the gradient $\nabla\mathcal{L}(\mathbf{w})$ "points to the direction" of $\mathbf{w}^*$. In particular we showed that we can pick the band with $a = \sin\theta(\mathbf{w},\mathbf{w}^*)$ and $b = -t\cos(\theta(\mathbf{w},\mathbf{w}^*))$. Unfortunately, we do not know either the value of the optimal threshold t or the angle between the guess and the optimal weight vector $\mathbf{w}^*$: $\theta(\mathbf{w},\mathbf{w}^*)$. It is not hard to obtain an estimate t' of the value of the threshold t from the noisy samples: we can show (see Claim C.9) that with $O(1/\epsilon)$ we can obtain a good estimate t'. In fact, we show that t' is the optimal threshold and only introduce $O(\epsilon)$ additional noise in the distribution $\mathcal{D}$. Therefore, to keep the presentation clean, in what follows we will assume that we know the value of the optimal threshold t. One could hope that we can also estimate the angle between $\mathbf{w}, \mathbf{w}^*$ from samples and assume that it is also known. It is unclear whether we can estimate the angle accurately enough: in fact we will show that we do not need to do so. In general, we will need a "robust" version of Lemma B.5 showing how close must be the threshold, and size values a, b to the "true" used in Lemma B.5 in order for the gradient $\nabla\mathcal{L}_k(\mathbf{w})$ to point to the right direction. We prove the following lemma that shows that it is ok to use the band $B_{a,b}$ with $a = \sin(\phi)$ and $b = t\cos(\phi)$ assuming that $|\phi - \theta(\mathbf{w},\mathbf{w}^*)| \le \min(\phi, 1/\log(1/\epsilon))$. The fact that we only have an (inverse) logarithmic tolerance (as opposed to requiring the difference to be $\mathrm{poly}(\epsilon)$ is crucial for obtaining the fast, i.e., in $\mathrm{poly}(\log(1/\epsilon)$ rounds, convergence of our Online Gradient Descent algorithm.

Lemma C.3 (Localized Update)**.** *Fix $\epsilon \in (0, 1/2)$ and a sufficiently large constant $C > 1$. Let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Denote by $\mathbf{w}^*$ the weight vector of an optimal halfspace, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x} + t) \ne y] \le \epsilon$. Fix $\phi \in [0, \pi/2]$ and a unit vector $\mathbf{w} \in \mathbb{R}^d$ and assume that $\min(e^{-t^2/2}\sin(\theta(\mathbf{w},\mathbf{w}^*)), e^{-t^2/2}/|t|) \ge C\epsilon$. Denote $B_\phi = \{z \in \mathbb{R} : |z - t\cos(\phi)| < \sin(\phi)\}$ and $\mathbf{g}(\mathbf{w},\phi) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \in B_\phi\}]$. Then, if $|\phi - \theta(\mathbf{w},\mathbf{w}^*)| \le \phi/(C\log(1/\epsilon))$, it holds that*

$$\mathbf{g}(\mathbf{w},\phi)\cdot\mathbf{w}^* \ge (\sin\theta)^2 e^{-t^2/2}\,.$$

Proof. Let $\theta = \theta(\mathbf{w},\mathbf{w}^*)$. First, we show that under the assumption that $|\phi - \theta(\mathbf{w},\mathbf{w}^*)| \le \phi/(C\log(1/\epsilon))$, it holds that $\max(|t\cos\phi - t\cos\theta|, |1 - (\sin\theta/\sin\phi)^2|) \le 1/C$. We have that $|t\cos\phi - t\cos\theta| \le |t||\phi - \theta| \le t/(C\log(1/\epsilon)) \le 1/C$ where we used that $|t| \le \sqrt{2\log(1/\epsilon)}$ from the assumptions, i.e., it holds that $\exp(-t^2/2)/|t| \ge C\epsilon$. Moreover, we need to show that $|1 - (\sin\theta/\sin\phi)^2| \le 1/C$. It suffices to show that $|1 - \sin\theta/\sin\phi| \le 1/(2C)$, because then $|1 - (\sin\theta/\sin\phi)^2| \le |1 + (\sin\theta/\sin\phi)|/(2C) \le 1/(2C) + 1/(2C)^2 \le 1/C$. From our assumptions, we have that $|\phi - \theta| \le \phi/(2C)$, therefore $|1 - (\sin\theta/\sin\phi)^2| \le 1/C$.

Denote $\bar{B} = [-t\cos\theta - \sin\theta, -t\cos\theta]$. We have that

$$\bar{p} = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(-t\cos\theta,(\sin\theta)^2)}[z \in \bar{B}] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,(\sin\theta)^2)}[|z| \le \sin\theta] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[|z| \le 1] \ge \frac{2}{3}\,.$$

Moreover, we have

$$\begin{aligned}\bar{q} &= \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[z \in \bar{B}] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[|z + t\cos\theta| \le \sin\theta] \le 2\sin\theta\max(e^{-(-t\cos\theta+\sin\theta)^2/2}, e^{-(-t\cos\theta-\sin\theta)^2/2})\\ &\le 2\sin\theta e^{-(t\cos\theta)^2/2+|t|\sin\theta+1/2} \le 4\sin\theta e^{-t^2/2}\, e^{(t\sin\theta)^2/2+|t|\sin\theta}\,.\end{aligned}$$

Let $a = \sin\phi$ and $b = -t\cos\phi$. We now show that under the assumptions that $\max(|b - t\cos\theta|, |1 - (\sin\theta/a)^2|) \le 1/C$ and $|a - \sin\theta| \le 1/(C\log(1/\epsilon))$ it holds that the the corresponding Gaussian integrals $p = \mathbf{Pr}_{z\sim\mathcal{N}(-t\cos\theta,(\sin\theta)^2)}[z \in B_\phi]$ and $q = \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \in B_\phi]$ are close to the values $\bar{p}$ and $\bar{q}$. In particular, we have the following claim.

Claim C.4. *It holds that $p \ge \bar{p}/2$ and $q \le 2\bar{q}$.*

Proof. We first prove that $p \geq \bar{p}/2$. We have that $\bar{p} = \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z| \leq 1]$ and $p = \mathbf{Pr}_{z\sim\mathcal{N}(-t\cos\theta - b, (\sin\theta/a)^2}[|z| \leq 1]$. For simplicity, denote $\mu = -t\cos\theta - b$ and $\sigma = \sin\theta/a$. The log-ratio of the densities of the standard normal and the normal $\mathcal{N}(\mu, \sigma^2)$ is equal to

$$\log\left(\frac{\mathcal{N}(0,1;z)}{\mathcal{N}(\mu,\sigma^2;z)}\right) = \log(\sigma) + (z-\mu)^2/(2\sigma^2) - z^2/2 = \log(\sigma) + \frac{z^2}{2}\left(\frac{1}{\sigma^2} - 1\right) - z\frac{\mu}{\sigma^2} + \frac{\mu^2}{2\sigma^2}\,.$$

Since we only care about the ratio in the interval $|z| \leq 1$ we can bound the above log-ratio in absolute value by $\log(\sigma) + |1 - 1/\sigma^2| + (|\mu| + |\mu|^2)/\sigma^2$. We see that when σ is sufficiently close to 1 (for concreteness $\sigma \in (1 - 1/100, 1 + 1/100)$ and $|\mu| \leq 1/100$ we obtain that for all $|z| \leq$ the (absolute value of the) above log-ratio is at most 0.05 which implies that p is well above $\bar{p}/2$.

We next bound the ratio of the probabilities $q, \bar{q}$. The main ingredient is the following lemma bounding the ratio of the probabilities of two different intervals under the Gaussian distribution. In particular, for the two intervals of the form $|x - b_1| \leq a_1$ and $|x - b_2| \leq a_2$ to have roughly the same Gaussian mass, it suffices that $a_1/a_2 = \Theta(1)$, $a_1 \leq 1$, $|b_1 - b_2| = O(1)$ and $|a_1^2 - a_2^2| \leq 1/(1 + |b_2|^2)$.

Lemma C.5 (Gaussian Intervals Ratio)**.** *Let $a_1, a_2 \in (0, \infty)$ and let $b_1, b_2 \in \mathbb{R}$. It holds*

$$e^{-c} \leq \frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_1| \leq a_1]}{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_2| \leq a_2]} \leq e^c,$$

where, $c = a_1^2/(|b_1 + b_2| + 1)|b_1 - b_2| + |\log(a_1/a_2)| + ((1 + |b_2|)^2/2)|a_1^2 - a_2^2|$.

Proof. It holds that

$$\frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_1| \leq a_1]}{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_2| \leq a_2]} = \frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_1| \leq a_1]}{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_2| \leq a_1]} \frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_2| \leq a_1]}{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_2| \leq a_2]}\,.$$

Therefore, it suffices to bound each of the ratios separately. For the first ratio, we have

$$\frac{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_1| \leq a_1]}{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[|z - b_2| \leq a_1]} = \frac{\mathbf{Pr}_{z\sim\mathcal{N}(-b_1, 1/a_1^2)}[|z| \leq 1]}{\mathbf{Pr}_{z\sim\mathcal{N}(-b_2, 1/a_1^2)}[|z| \leq 1]}$$

We have that the log-ratio of the densities of $\mathcal{N}(-b_1, 1/a_1^2)$ and $\mathcal{N}(-b_2, 1/a_1^2)$ is equal to

$$\log\left(\frac{\mathcal{N}(-b_1, 1/a_1^2; x)}{\mathcal{N}(-b_2, 1/a_1^2; x)}\right) = -(x + b_1)^2 a_1^2/2 + (x + b_2)^2 a_1^2/2 = a_1^2/2(x(b_2 - b_1) + (b_2^2 - b_1^2),$$

which, in absolute value, is less than or equal to $a_1^2/2(|b_1 + b_2| + 1)|b_1 - b_2|$, where we used the fact that we only need to bound the ratio of the densities in the interval $|x| \leq 1$. Similarly, we bound the log-ratio of densities of the second ratio

$$\log\left(\frac{\mathcal{N}(-b_2, 1/a_1^2; x)}{\mathcal{N}(-b_2, 1/a_2^2; x)}\right) = \log(\frac{a_1}{a_2}) - (x - b_2)^2 a_1^2/2 + (x - b_2)^2 a_2^2/2 = \log(\frac{a_1}{a_2}) + -((x - b_2)^2/2)(a_2^2 - a_1^2),$$

which, in absolute value is at most $|\log(a_1/a_2)| + ((1 + |b_2|)^2/2)|a_1^2 - a_2^2|$.

□

From the assumption that $|t| \leq \sqrt{2\log(1/\epsilon)}$ we have that $|t\cos\theta| \leq \sqrt{2\log(1/\epsilon)}$, and therefore, using Lemma C.5, we obtain that when $|b + t\cos\theta|$, $|1 - (\sin\theta/a)^2|$ are sufficiently small universal constants, and $|\sin\theta - a|$ is a sufficiently small constant multiple of $1/\log(1/\epsilon)$, it holds that $\bar{q} \leq 2q$.

□

Therefore, we have that $\log(\bar{q}/\epsilon+1) \leq \log(\sin\theta e^{-t^2/2}/\epsilon) + (t\sin\theta)^2/2 + |t|\sin\theta + 3$. We next denote $\bar{\mathbf{g}} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}\mathbb{1}_{\bar{B}}(\mathbf{w}\cdot\mathbf{x})]$. Using Lemma B.7, and the inequality $\sqrt{a}+\sqrt{b} \leq \sqrt{2}\sqrt{a+b}$, we have that

$$\begin{aligned}\frac{1}{\sin\theta}\mathbf{g}(\mathbf{w},\phi)\cdot\mathbf{w}^* &\geq \sqrt{2\frac{2}{\pi}}\left(\sin\theta e^{-\frac{t^2}{2}}\bar{p} - 18\sqrt{2}\epsilon\sqrt{\log(\bar{q}/\epsilon+1)}\right)\\ &\geq \frac{2\sqrt{2}}{3\sqrt{\pi}}\left(\sin\theta e^{-\frac{t^2}{2}} - 60\epsilon\big(\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+3} + \sqrt{(t\sin\theta)^2+|t|\sin\theta}\big)\right)\,,\end{aligned}$$

Observe now that since, by the assumptions of Lemma C.3, it holds that $\sin\theta e^{-t^2/2}/\epsilon$ is greater than some sufficiently large absolute constant $C > 1$, it suffices to show that it is larger than each of the "noise" terms: $\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+3}$, $\sqrt{(t\sin\theta)^2+|t|\sin\theta}$, separately. We will first prove that the term $\sin\theta e^{-t^2/2}/3$ is greater than $20\epsilon\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+3}$. Observe that since $\sin\theta e^{-t^2/2} \geq 1$ it suffices to show that it is larger than $20\sqrt{3}\epsilon\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+1}$. From Claim B.8, we obtain that when $\sin\theta e^{-t^2/2}/\epsilon \geq (60\cdot 3\sqrt{3})^2$ it holds that $\sin\theta e^{-t^2/2}/3 \geq 60\sqrt{3}\epsilon\sqrt{\log(\sin\theta e^{-t^2/2}/\epsilon)+1}$. We next show that $\sin\theta e^{-t^2/2}/3 \geq 60\epsilon\sqrt{(t\sin\theta)^2+|t|\sin\theta}$. We distinguish two cases. First, we handle the case $|t|\sin\theta \leq 1$. In this case, we have to show that $\sin\theta e^{-t^2/2}/3 \geq 20\sqrt{6}\epsilon$ which holds directly from assumptions. When $|t|\sin\theta \geq 1$ we have that $\sqrt{(t\sin\theta)^2+|t|\sin\theta} \leq \sqrt{2}|t|\sin\theta$. Therefore, in order for $\sin\theta e^{-t^2/2}/3$ to be larger than $20\sqrt{2}\epsilon|t|\sin\theta$ we need $e^{-t^2/2}/|t| \geq 3\cdot 20\sqrt{2}\epsilon$, which holds from the assumption that $e^{-t^2/2}/(\epsilon|t|)$ is greater than an absolute constant.

□

We now have shown that there exists a choice for the parameters a, b of the ramp objective $\mathcal{L}(\mathbf{w})$ so that its gradient point to the right direction. Given a current guess $\mathbf{w}$ we still have the issue of knowing the angle between $\mathbf{w}$ and $\mathbf{w}^*$. In the next lemma we show that we do not need to know the value of the angle but only an upper bound, i.e., we know that $\theta(\mathbf{w},\mathbf{w}^*) \in [0,\phi]$. We show that, with an appropriate step size, the online gradient descent update of Algorithm 1 is a contraction map in the sense that, even though an update may make the current angle between the guess $\mathbf{w},\mathbf{w}^*$ worse, we can show that the new angle belongs in a smaller interval $[0,\phi']$ that is significantly smaller than our initial interval $[0,\phi]$.

Lemma C.6 (Angle Contractive Map)**.** *Fix a unit vector $\mathbf{v}^* \in \mathbb{R}^d$, $\beta,\gamma,\kappa \in (0,1)$ and a vector field $\mathbf{g}: \mathbb{R}^d \times \mathbb{R} \mapsto \mathbb{R}^d$ such that for any vector $\mathbf{u}$ and $\phi \in [0,\pi/2]$, it holds that $\|\mathbf{g}(\mathbf{u},\phi)\|_2 \leq \kappa$, $\mathbf{g}(\mathbf{u},\phi)\cdot\mathbf{u} = 0$, and, if $|\phi-\theta(\mathbf{u},\mathbf{v}^*)| \leq \gamma\sin\phi$, and $\sin(\theta(\mathbf{u},\mathbf{v}^*)) \geq \beta > 0$ then $\mathbf{g}(\mathbf{u},\phi)\cdot\mathbf{v}^* \geq \rho\kappa\sin(\theta(\mathbf{u},\mathbf{v}^*)) > 0$. Fix $\phi \in [0,\pi/2]$ with $\sin(\phi) \geq \beta > 0$. Set $\mathbf{v}$ to be any unit vector in $\mathbb{R}^d$ with $\theta(\mathbf{v},\mathbf{v}^*) \leq \phi$ and consider the normalized gradient update rule*

$$\mathbf{v}' = \frac{\mathbf{v}+\lambda\mathbf{g}(\mathbf{v},\phi)}{\|\mathbf{v}+\lambda\mathbf{g}(\mathbf{v},\phi)\|_2}\,,$$

with $\lambda = \rho\gamma\sin\phi/(4\kappa)$. Then, it holds that $\theta(\mathbf{v}',\mathbf{v}^*) \leq \phi'$, where, $\phi' = \max(\phi(1-\rho^2\gamma^2/64),\beta)$.

Proof. First, we start by proving the following claim.

Claim C.7 (Correlation Improvement)**.** *For unit vectors $\mathbf{v}^*,\mathbf{v} \in \mathbb{R}^d$, let $\mathbf{u} \in \mathbb{R}^d$ such that $\mathbf{u}\cdot\mathbf{v}^* \geq c$, $\mathbf{u}\cdot\mathbf{v} = 0$, and $\|\mathbf{u}\|_2 \leq 1$, with $c > 0$. Then, for $\mathbf{v}' = \frac{\mathbf{v}+\lambda\mathbf{u}}{\|\mathbf{v}+\lambda\mathbf{u}\|_2}$, with $\lambda \leq c/2$, we have that $\mathbf{v}'\cdot\mathbf{v}^* \geq \mathbf{v}\cdot\mathbf{v}^* + \lambda c/8$.*

Proof. We will show that $\mathbf{v}'\cdot\mathbf{v}^* = \cos\theta' \geq \cos\theta + \lambda^2/2$, where $\cos\theta = \mathbf{v}\cdot\mathbf{v}^*$. We have that

$$\|\mathbf{v}+\lambda\mathbf{u}\|_2 = \sqrt{1+\lambda^2\|\mathbf{u}\|_2^2} \leq 1+\lambda^2\|\mathbf{u}\|_2^2\,, \tag{5}$$

where we used that $\sqrt{1+a} \leq 1+a$, for $a > 0$. Using the update rule, we have

$$\mathbf{v}'\cdot\mathbf{v}^* = \mathbf{v}'\cdot(\mathbf{v}^*)^{\perp_\mathbf{v}}\sin\theta + \mathbf{v}'\cdot\mathbf{v}\cos\theta = \frac{\lambda\mathbf{u}\cdot(\mathbf{v}^*)^{\perp_\mathbf{v}}}{\|\mathbf{v}+\lambda\mathbf{u}\|_2}\sin\theta + \frac{(\mathbf{v}+\lambda\mathbf{u})\cdot\mathbf{v}}{\|\mathbf{v}+\lambda\mathbf{u}\|_2}\cos\theta\,.$$

Now using Equation (5), we get

$$\mathbf{v}'\cdot\mathbf{v}^* \geq \frac{\lambda\mathbf{u}\cdot(\mathbf{v}^*)^{\perp_\mathbf{v}}}{1+\lambda^2\|\mathbf{u}\|_2^2}\sin\theta + \frac{\cos\theta}{1+\lambda^2\|\mathbf{u}\|_2^2} = \cos\theta + \frac{\lambda\mathbf{u}\cdot(\mathbf{v}^*)^{\perp_\mathbf{v}}}{1+\lambda^2\|\mathbf{u}\|_2^2}\sin\theta + \frac{-\lambda^2\|\mathbf{u}\|_2^2\cos\theta}{1+\lambda^2\|\mathbf{u}\|_2^2}\,.$$

Then, using that $\mathbf{u} \cdot \mathbf{v}^* = \mathbf{u} \cdot (\mathbf{v}^*)^{\perp_{\mathbf{v}}} \sin\theta$, we have that $\mathbf{u} \cdot (\mathbf{v}^*)^{\perp_{\mathbf{v}}} \geq \frac{c}{\sin\theta}$, thus

$$\mathbf{v}' \cdot \mathbf{v}^* \geq \cos\theta + \frac{\lambda c - \lambda^2 \|\mathbf{u}\|_2^2}{1 + \lambda^2 \|\mathbf{u}\|_2^2} \geq \cos\theta + \frac{\lambda c - \lambda^2}{1 + \lambda^2 \|\mathbf{u}\|_2^2} = \cos\theta + \frac{1}{4}\frac{\lambda c}{1 + \lambda^2 \|\mathbf{u}\|_2^2} ,$$

where in the first inequality we used that $\|\mathbf{u}\|_2 \leq 1$ and in the second that for $\lambda \leq c/2$ it holds $c - \lambda \geq c/2$. Finally, we have that

$$\cos\theta' = \mathbf{v}' \cdot \mathbf{v}^* \geq \cos\theta + \frac{1}{4}\frac{c\lambda}{1 + \lambda^2 \|\mathbf{u}\|_2^2} \geq \cos\theta + \frac{1}{8}c\lambda .$$

This completes the proof. □

We shall distinguish two cases. In the first case we assume that $|\phi - \theta(\mathbf{v}, \mathbf{v}^*)| \leq \gamma \sin\phi$ and that means that it holds $\mathbf{g}(\mathbf{v}, \phi) \cdot \mathbf{v}^* \geq \rho \sin(\theta(\mathbf{v}, \mathbf{v}^*))/\kappa$; then from Claim C.7, we have that $\mathbf{v}' \cdot \mathbf{v}^* \geq \mathbf{v} \cdot \mathbf{v}^* + \lambda\rho \sin(\theta(\mathbf{v}, \mathbf{v}^*))/(8\kappa) \geq \mathbf{v} \cdot \mathbf{v}^* + \rho^2\gamma^2\phi^2/64$, where we used that $\lambda = \rho\gamma \sin\phi/(4\kappa) \geq \rho\gamma\phi/(8\kappa)$, because $\sin x \geq x/2$ for $\pi/2 \geq x \geq 0$. Hence, we have that $\cos(\theta(\mathbf{v}', \mathbf{v}^*)) \geq \cos(\theta(\mathbf{v}, \mathbf{v}^*)) + \rho^2\gamma^2\phi^2/64$ and note that because $\cos(t)$ is decreasing in $[0, \pi]$, it holds that $\theta(\mathbf{v}, \mathbf{v}^*) \geq \theta(\mathbf{v}', \mathbf{v}^*)$. Moreover, using the trigonometric identity $\cos x - \cos y = 2\sin((x+y)/2)\sin((y-x)/2)$ and that $\sin x \leq x$ for $x > 0$, we get that

$$\begin{aligned}\cos(\theta(\mathbf{v}', \mathbf{v}^*)) - \cos(\theta(\mathbf{v}, \mathbf{v}^*)) &= 2\sin((\theta(\mathbf{v}', \mathbf{v}^*) + \theta(\mathbf{v}, \mathbf{v}^*))/2)\sin((\theta(\mathbf{v}, \mathbf{v}^*) - \theta(\mathbf{v}', \mathbf{v}^*))/2) \\ &\leq \theta(\mathbf{v}, \mathbf{v}^*)^2/2 - \theta(\mathbf{v}', \mathbf{v}^*)^2/2 ,\end{aligned}$$

hence,

$$\theta(\mathbf{v}, \mathbf{v}^*)^2 - \theta(\mathbf{v}', \mathbf{v}^*)^2 \geq \rho^2\gamma^2\phi^2/32 ,$$

and using that $\theta(\mathbf{v}, \mathbf{v}^*) \leq \phi$, we get that

$$\phi^2(1 - \rho^2\gamma^2/32) \geq \theta(\mathbf{v}', \mathbf{v}^*)^2 ,$$

which completes the proof for this case, since we have shown that $\phi' = \phi(1 - \rho^2\gamma^2/64) \geq \phi(1 - \rho^2\gamma^2/32)^{1/2} \geq \theta(\mathbf{v}', \mathbf{v}^*)$.

We now assume that the true current angle $\theta(\mathbf{v}, \mathbf{v}^*)$ is far from our current upper bound ϕ, i.e., $|\phi - \theta(\mathbf{v}, \mathbf{v}^*)| \geq \gamma \sin\theta$. In this case, we will potentially do a step in the wrong direction, i.e., the angle between $\mathbf{v}'$ and $\mathbf{v}^*$ will become worse than before but still not worse than our new upper bound for the angle ϕ'. By Cauchy-Schwarz, and the fact that $\|\mathbf{g}(\mathbf{v}, \phi)\|_2 \leq \kappa$, we obtain $\|\mathbf{v}' - \mathbf{v}\|_2 \leq 2\lambda\kappa$. Equivalently, we have that $|\cos(\theta(\mathbf{v}', \mathbf{v}^*)) - \cos(\theta(\mathbf{v}, \mathbf{v}^*))| \leq 2\lambda\kappa$. Using the fact that $t \mapsto \cos(t)$ is 1-Lipschitz we obtain that $\theta(\mathbf{v}', \mathbf{v}^*) \leq \theta(\mathbf{v}, \mathbf{v}^*) + 2\kappa\lambda$. Since the initial angle was far from the upper bound ϕ, $\phi - \theta(\mathbf{v}, \mathbf{v}^*) \geq \gamma \sin\phi$, we have that

$$\phi' - \theta(\mathbf{v}', \mathbf{v}^*) \geq (\phi' - \phi) + (\phi - \theta(\mathbf{v}, \mathbf{v}^*)) - 2\lambda\kappa \geq \gamma \sin\theta > 0 .$$

Therefore, we conclude that, in both cases, the new angle $\theta(\mathbf{v}', \mathbf{v}^*)$ belongs in the interval $[0, \phi']$. □

Corollary C.8. *Fix unit vectors $\mathbf{v}^*, \mathbf{v}^{(0)} \in \mathbb{R}^d$ with $\theta(\mathbf{v}^{(0)}, \mathbf{v}^*) \leq \pi/2$, parameters $\beta, \gamma \in (0, 1)$, and function $t : \mathbb{R} \mapsto (0, 1]$. Moreover, fix a vector field $\mathbf{g} : \mathbb{R}^d \times \mathbb{R} \mapsto \mathbb{R}^d$ such that for any vector $\mathbf{u}$ and $\phi \in [0, \pi/2]$, it holds that $\|\mathbf{g}(\mathbf{u}, \phi)\|_2 \leq t(\phi) \leq 1$, $\mathbf{g}(\mathbf{u}, \phi) \cdot \mathbf{u} = 0$, and, if $|\phi - \theta(\mathbf{u}, \mathbf{v}^*)| \leq \gamma \sin\phi$, and $\sin(\theta(\mathbf{u}, \mathbf{v}^*)) \geq \beta > 0$ then $(1/t(\phi))(\mathbf{g}(\mathbf{u}, \phi) \cdot \mathbf{v}^*) \geq \rho \sin(\theta(\mathbf{u}, \mathbf{v}^*)) > 0$, where $\rho > 0$. Fix sequence $\alpha_i = (\pi/2)(1 - \rho^2\gamma^2/64)^i$ and let $\lambda_i = \rho \sin\alpha_i\gamma/(4t(\alpha_i))$ and consider the normalized gradient update rule*

$$\mathbf{v}^{(i+1)} = \frac{\mathbf{v}^{(i)} + \lambda_i \mathbf{g}(\mathbf{v}^{(i)}, \alpha_i)}{\left\|\mathbf{v}^{(i)} + \lambda_i \mathbf{g}(\mathbf{v}^{(i)}, \alpha_i)\right\|_2} .$$

Then, after $T = O(\log(1/\beta)/(\gamma\rho)^2)$ steps, it holds that $\theta(\mathbf{v}^{(T)}, \mathbf{v}^*) \leq 2\beta$.

Proof. We show that for each $i \in \mathbb{N}$ with $i \leq O(\log(1/\beta)/(\gamma\rho)^2)$, it holds that $\theta(\mathbf{v}^{(i)}, \mathbf{v}^*) \leq \alpha_i$. For the base case $k = 0$, it holds trivially from the assumptions. Assume that for $k = i$ it holds $\theta(\mathbf{v}^{(k)}, \mathbf{v}^*) \leq \alpha_k$, we will show that if $\alpha_{k+1} \geq \beta$, then $\theta(\mathbf{v}^{(k+1)}, \mathbf{v}^*) \leq \alpha_{k+1}$. From Lemma C.6, we have for $\phi = \alpha_k$ that one gradient update step will give a $\mathbf{v}^{k+1}$ such that $\theta(\mathbf{v}^{(k+1)}, \mathbf{v}^*) \leq \phi(1 - \rho^2\gamma^2/64) = \alpha_{k+1}$; and therefore from mathematical induction we get that $\theta(\mathbf{v}^{(i)}, \mathbf{v}^*) \leq \alpha_i$ for each $i \in \mathbb{N}$ with $a_i \geq 2\beta$. To find the maximum number of steps, note that $a_T = (\pi/2)(1 - \rho^2\gamma^2/64)^T \leq (\pi/2)\exp(-\rho^2\gamma^2 T/64)$. Hence, for $T = O(\log(1/\beta)/(\gamma\rho)^2)$, we get that $a_T \leq 2\beta$. □

C.1. The Proof of Theorem 1.2

We first show that the simple empirical threshold estimate obtained in the first step of Algorithm 2 is sufficiently good.

Claim C.9 (Threshold estimation)**.** *Fix $\epsilon, \delta' \in (0, 1/2)$ and let $\mathcal{D}$ be an ϵ-corrupted distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal and let $\mathbf{w}^*, t$ be a unit vector and a threshold such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t) \neq y] \leq \epsilon$. Denote $\widehat{\mathcal{D}}$ the empirical distribution with $N = O(\log(1/\delta')/\epsilon^2)$ samples and let $t' = \Phi^{-1}\left((1 - \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[y])/2\right)$ be the empirical estimate of t used in Algorithm 2. It holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t') \neq y] \leq O(\epsilon)$, with probability at least $1 - \delta'$.*

Proof. First, note that for any $t' \in \mathbb{R}$ with $|\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t'] - \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t]| \leq \epsilon'$, for some $\epsilon' = \Theta(\epsilon)$, we have that

$$\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t')] = |\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t'] - \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t]| = \epsilon' \,.$$

Hence, $\mathbf{Pr}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t') \neq y] \leq \epsilon + \epsilon' = \Theta(\epsilon)$. Therefore, by assuming that the unknown threshold is any t' with $|\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t'] - \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t]| = \Theta(\epsilon)$, we introduce $O(\epsilon)$ noise.

To find such a t', note that for the unknown threshold t, it holds that $|\mathbf{E}[y] - \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \geq t] + \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t]| \leq \epsilon$ which is equivalent to $|(\mathbf{E}[y] - 1)/2 + \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t]| \leq \epsilon/2$. Moreover, from Hoeffding inequality, we have that with $O(\log(1/\delta')/\epsilon^2)$ samples with probability $1 - \delta'$, it holds that $|\mathbf{E}_{y\sim\widehat{\mathcal{D}}_y}[y] - \mathbf{E}[y]| \leq \epsilon$, we call this event A_0. Therefore, $|(\mathbf{E}_{y\sim\widehat{\mathcal{D}}_y}[y] - 1)/2 + \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t]| \leq \epsilon$. For $t' = \Phi^{-1}((1 - \mathbf{E}_{y\sim\widehat{\mathcal{D}}_y}[y])/2)$, we get $|\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t'] - \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \leq t]| \leq \epsilon$. Hence, for this choice of t', we have that $\mathbf{Pr}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t') \neq y] = O(\epsilon)$. □

Therefore, from Claim C.9, using t' as the optimal threshold and only introduce $O(\epsilon)$ additional noise in the distribution $\mathcal{D}$. Therefore, to keep the presentation clean, in what follows we will assume that we know the value of the optimal threshold t.

To prove Theorem 1.2, we need to consider several cases. The first case is when $\exp(-t^2/2)/|t| \leq C\epsilon$. In this case, any unit vector $\mathbf{w} \in \mathbb{R}^d$ with the correct threshold t will have error $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t)) \neq y] \leq 3C\epsilon$. To show this, we need the following fact that bounds from below the tail of the standard normal distribution.

Fact C.10 (Komatsu's Inequality, see, e.g., (Kouba, 2006))**.** *for any $t \geq 0$, it holds that*

$$\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \geq t] \leq \frac{4\exp(-t^2/2)}{3t + \sqrt{t^2 + 8}} \,.$$

Therefore, if $\exp(-t^2/2)/t \leq C\epsilon$, then $\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \geq t] \leq C\epsilon$, hence $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t)) \neq \mathrm{sign}(t)] \leq C\epsilon$ which from triangle inequality gives

$$\begin{aligned}
\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] &\leq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t) \neq y] + \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t)] \\
&\leq \epsilon + \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t) \neq \mathrm{sign}(t)] + \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq \mathrm{sign}(t)] \\
&\leq 3C\epsilon \,.
\end{aligned}$$

Next, we consider the case where $\exp(-t^2/2)/|t| \geq C\epsilon$. For simplicity, for the rest of the proof, we will assume that we know the optimal threshold t and use the ϵ for the new noise rate which is c times more than the previous one, where $c > 0$ is some absolute constant.

We are going to show that after T gradient steps, we will get a $\mathbf{w}^{(T)}$ such that $\theta(\mathbf{w}^{(T)}, \mathbf{w}^*) \leq O(\exp(t^2/2)\epsilon)$. Then using the following fact, which connects the disagreement of two hypotheses with the distance between these vectors, we will show that the current hypothesis gets small error.

Fact C.11 ((Diakonikolas et al., 2018))**.** *Fix unit vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ and $t \in \mathbb{R}$, it holds*

$$\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{sign}(\mathbf{w}_1 \cdot \mathbf{x} + t) \neq \mathrm{sign}(\mathbf{w}_2 \cdot \mathbf{x} + t)] = \frac{1}{\pi}\int_0^{\theta(\mathbf{w}_1,\mathbf{w}_2)} e^{-t^2/(1+\cos\phi)}\mathrm{d}\phi \leq \frac{\theta(\mathbf{w}_1, \mathbf{w}_2)}{\pi}\exp(-t^2/2) \,.$$

Note that using the fact above, when $\theta(\mathbf{w}^{(T)}, \mathbf{w}^*) \leq O(\exp(t^2/2)\epsilon)$, we get that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{sign}(\mathbf{w}^{(T)} \cdot \mathbf{x} + t) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t)] = O(\epsilon)$. From Claim C.2, we have that $\nabla_{\mathbf{w}} \mathcal{L}_k(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x} - b_k| \leq a_k\}]$.

Next, we show that when $\mathbf{w}^{(k)}$, a_k and b_k satisfy certain conditions, we have that $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}^{(k)}) \cdot \mathbf{w}^* \geq c' \sin^2\theta \exp(-t^2/2)$, where $\theta = \theta(\mathbf{w}^*, \mathbf{w}^{(k)})$. Let $\phi \in (0,\pi)$ and denote $\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y) = y(\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{x})\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x} - t\cos(\phi)| \leq \sin(\phi)\}$ and $B_\phi = \{z \in \mathbb{R} : |z - t\cos(\phi)| \leq \sin\phi\}$.

We will show that if N is large enough, then for any unit vector $\mathbf{u} \in \mathbb{R}^d$, we will have that $\mathbf{u} \cdot \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]$ is close to $\mathbf{u}\cdot\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]$. The proof of Lemma C.12 can be found on Appendix C.2.

Lemma C.12. *Let $q = \mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x} \in B_\phi]$. Assuming that $N \geq O\left(\frac{d\log(1/\delta)q}{\epsilon'^2} + \frac{\log(d/\delta)}{q^2}\right)$, we have with probability at least $1-\delta$ that for any unit vector $\mathbf{u} \in \mathbb{R}^d$, it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{u}\cdot\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq (1/2)\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{u}\cdot\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] - \epsilon' \,,$$

and that $\|\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2 \leq 2\|\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{u}\cdot\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2 + \epsilon'$.

Assuming that $\phi \geq C\epsilon\exp(t^2/2)$, we have that $\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x} \in B_\phi] \geq \epsilon$, to see this, observe that $\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x} \in B_\phi] \geq 1/(\sqrt{2\pi})\exp(-(t\cos\theta + \sin\theta)^2/2)\sin\theta \geq \epsilon\exp(t^2\sin^2\theta/2) \geq \epsilon$. Therefore, from Lemma C.12 given $N \geq O(\frac{d\log(1/\delta')}{\epsilon^2})$, we have with probability at least $1-\delta'$ that for any unit vector $\mathbf{u} \in \mathbb{R}^d$, it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{u}\cdot\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq (1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{u}\cdot\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] - \epsilon$ and $\|\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2 \leq 2$. We denote A_k this event and note that this event has probability at least $1-\delta'$, i.e., $\mathbf{Pr}[A_k] \geq 1-\delta'$.

Notice that from Lemma C.3, if $\sin\theta\exp(-t^2/2) \geq C\epsilon$ and $|\theta - \phi| \leq (1/C)\min(\sin\phi, 1/\log(1/\epsilon))$, where $C > 0$ is a sufficiently large constant, we have that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{w}^*\cdot\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq (\sin\theta)^2\exp(-t^2/2)$ and hence, for the unit vector $\mathbf{v} = (\mathbf{w}^*)^{\perp\mathbf{w}}/\|(\mathbf{w}^*)^{\perp\mathbf{w}}\|_2$, we have that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{v}\cdot\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq \sin\theta\exp(-t^2/2)$, where we used the fact that $\|(\mathbf{w}^*)^{\perp\mathbf{w}}\|_2 = \sin\theta$.

Therefore, conditioning on the event A_k, we have that $\mathbf{v}\cdot\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq \sin\theta\exp(-t^2/2) - \epsilon \geq \sin\theta\exp(-t^2/2)/2$, where used again the fact $\epsilon \leq C\sin\theta\exp(-t^2/2)/2$. Hence, we have that

$$\mathbf{w}^* \cdot \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq \sin^2\theta\exp(-t^2/2)/2 \,.$$

Moreover, we have that $\|\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2 \leq 2\|\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2 + \epsilon$, and from Lemma B.2 we have that $\|\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2 = O(\sin\theta\exp(-t^2/2)\sqrt{\log(1/\epsilon)})$, which gives $\|\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2 = O(\sin\theta\exp(-t^2/2)\sqrt{\log(1/\epsilon)}) = O(\sin\phi\exp(-t^2/2)\sqrt{\log(1/\epsilon)}) = \kappa(\phi)$, where we used that $|\sin\theta - \sin\phi| \leq \sin\phi/C$, from the assumptions. Hence, we have that

$$\frac{1}{\kappa(\phi)}\mathbf{w}^* \cdot \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq c''\sin\theta/\sqrt{\log(1/\epsilon)} \,,$$

where c'' is a sufficiently small constant. Therefore, assuming that $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq \pi/2$ by applying Corollary C.8 with parameters $\beta = C\exp(t^2/2)\epsilon$, $\gamma = (1/C\log(1/\epsilon))$, $\rho = c''/\log(1/\epsilon)$, and $\kappa(\phi) = \min(C\sin\phi\exp(-t^2/2)\sqrt{\log(1/\epsilon)}, 1)$, we get that after $T = O(\log^4(1/\epsilon))$ update steps, we have that conditioning on the events $A_0, A_1, \ldots, A_T$, we get $\theta(\mathbf{w}^{(T)}, \mathbf{w}^*) \leq \exp(t^2/2)\epsilon/c'$ and from Fact C.11, we get that

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{sign}(\mathbf{w}^{(T)}\cdot\mathbf{x} + t) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x} + t)] = O(\epsilon) \,.$$

Moreover, the algorithm fails if one of the events A_k does not happen, which from union bound has probability at most $T\delta'$ and by setting $\delta' = \delta/T = O(\delta/\log^3(1/\epsilon))$, we get overall sample complexity $\tilde{O}(\frac{d\log(1/\delta)}{\epsilon^2})$. To complete the proof, we need to show that $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq \pi/2$. Let $\mathbf{w}^{(0)} = \mathbf{E}_{(\mathbf{x},y)\sim}[\mathbf{x}y]$ and we show that $\mathbf{w}^*\cdot\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x}y] \geq 0$. Let $S(\mathbf{x},y)$ be

the event that a sample $(\mathbf{x}, y)$ is corrupted. We have that

$$\begin{aligned}\mathbf{w}^* \cdot \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x}y] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}^* \cdot \mathbf{x}\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t)] - 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{w}^* \cdot \mathbf{x}|S(\mathbf{x}, y)] \\ &= \int_{-|t|}^{\infty} \frac{2t}{\sqrt{2\pi}} e^{-z^2/2} \mathrm{d}z - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{w}^* \cdot \mathbf{x}|S(\mathbf{x}, y)] \\ &= \frac{2}{\sqrt{2\pi}} e^{-t^2/2} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{w}^* \cdot \mathbf{x}|S(\mathbf{x}, y)] \\ &\geq \frac{2}{\sqrt{2\pi}} e^{-t^2/2} - 2\sqrt{e}\epsilon\sqrt{\log(1/\epsilon + 1)}\,,\end{aligned}$$

where in the first inequality we used Lemma B.4. To show that this is positive, note that if $|t| \leq \sqrt{\log(1/(2C\epsilon\sqrt{\log(1/\epsilon)}))}$, then $\frac{2}{\sqrt{2\pi}} e^{-t^2/2} - 2\sqrt{e}\epsilon\sqrt{\log(1/\epsilon + 1)} > 0$. For the other case, note that from our assumptions, we have that $\exp(-t^2/2)/|t| \geq C\epsilon$, and for $t = \sqrt{\log(1/(2C\epsilon\sqrt{\log(1/\epsilon)}))}$, we get

$$\frac{2C\epsilon\sqrt{\log(1/\epsilon)}}{\sqrt{\log(1/(2C\epsilon\sqrt{\log(1/\epsilon)}))}} = \frac{2C\epsilon\sqrt{\log(1/\epsilon)}}{\sqrt{\log(1/(2C\epsilon)) - \sqrt{\log(1/\epsilon)}}} \leq \frac{2C\epsilon\sqrt{\log(1/\epsilon)}}{\sqrt{\frac{1}{2}\log(1/(2C\epsilon))}} \leq C\epsilon\,,$$

therefore, $|t| \leq \sqrt{\log(1/(2C\epsilon\sqrt{\log(1/\epsilon)}))}$ and hence, we have $\mathbf{w}^* \cdot \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x}y] > 0$. Hence, similar with Claim C.13, with $O(d\log(1/\delta))$ samples, we have that $\mathbf{w}^* \cdot \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{x}y] > 0$, therefore, $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq \pi/2$.

C.2. Proof of Lemma C.12

First we prove the following claim:

Claim C.13 (Uniform Convergence of $\mathbf{g}$)**.** *Fix $\epsilon, \delta \in (0, 1/2)$. Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Fix unit vector $\mathbf{w} \in \mathbb{R}^d$ and let $B = \{\mathbf{x} \in \mathbb{R}^d : t_1 \leq \mathbf{w} \cdot \mathbf{x} \leq t_2\}$ with $t_1, t_2 \in \mathbb{R}$. Denote $\mathbf{g}(\mathbf{x}, y) = \mathrm{proj}_{\mathbf{w}^\perp}(\mathbf{x})y$, and $\mathcal{D}_B$ be the distribution $\mathcal{D}$ conditioned on B. Moreover, let $\widehat{\mathcal{D}}$ be the empirical distribution of $\mathcal{D}_B$ with $N > 0$ samples. Then, if $N \geq O(\frac{d\log(1/\delta)}{\epsilon^2})$ from $\mathcal{D}_B$, with probability at least $1 - \delta$, it holds that $\| \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}_B}[\mathbf{g}(\mathbf{x}, y)] - \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_N}[\mathbf{g}(\mathbf{x}, y)]\|_2 \leq \epsilon$.*

Proof. First, we show that the random vector $\mathbf{g}$ is subgaussian with parameter 1. We remind that we say that a random variable X is subgaussian with parameter β if $\beta = \inf\{z > 0 : \mathbf{E}[\exp(X^2/z^2)] \leq 2\}$ and we say that a random vector $\mathbf{X}$ is β subgaussian if for any unit vector $\mathbf{v} \in \mathbb{R}^d$, the random variable $\mathbf{v} \cdot \mathbf{X}$ is subguassian with parameter β. Let $\mathbf{u} \in \mathbb{R}^d$ be any unit vector. We write $\mathbf{u} = a\mathbf{w} + b\mathbf{u}^{\perp\mathbf{w}}$ and notice that $\mathbf{g} \cdot \mathbf{u} = b\,\mathbf{g} \cdot \mathbf{u}^{\perp\mathbf{w}}$. We have that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}_B}[\exp((\mathbf{g}(\mathbf{x}, y) \cdot \mathbf{u})^2/z^2)] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\exp((y\mathbf{x} \cdot \mathbf{u}^{\perp\mathbf{w}})^2/z^2)] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\exp((\mathbf{x} \cdot \mathbf{u}^{\perp\mathbf{w}})^2/z^2)]\,,$$

where in the first inequality we used that $\mathbf{g}$ is perpendicular to $\mathbf{w}$. Moreover, because $\mathcal{D}_\mathbf{x}$ is standard d-dimensional normal, we have that for $z = 1$, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\exp((\mathbf{x} \cdot \mathbf{u}^{\perp\mathbf{w}})^2/z^2)] \leq 2$, therefore $\mathbf{g}$ is subgaussian with parameter 1. Next, we make use of the following fact which shows that the norm of $\mathbf{g}$ is concentrated well enough.

Fact C.14 (Lemma 1 of (Jin et al., 2019))**.** *If a random vector $\mathbf{x}$ is subgaussian with parameter β, then there exists an absolute constant $c > 0$ such that*

$$\mathbf{Pr}[\|\mathbf{x} - \mathbf{E}[\mathbf{x}]\|_2 \geq t] \leq 2\exp(-t^2/(c\beta^2 d))\,.$$

Using a simple application of the above, we get that with $N \geq O(d\log(1/\delta)/\epsilon^2)$ samples from $\mathcal{D}_B$, we have with probability at least $1 - \delta$ that $\| \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}_B}[\mathbf{g}(\mathbf{x}, y)] - \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_N}[\mathbf{g}(\mathbf{x}, y)]\|_2 \leq \epsilon$, which completes the proof of Claim C.13.

□

From Claim C.13, we have that with $O(d\log(1/\delta)/\epsilon'^2)$ samples from the region B, we have that with probability at least $1-\delta$ it holds for any unit vector $\mathbf{u}\in\mathbb{R}^d$ that $\mathbf{u}\cdot\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)\mid \mathbf{w}\cdot\mathbf{x}\in B_\phi] \geq \mathbf{u}\cdot\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)\mid \mathbf{w}\cdot\mathbf{x}\in B_\phi]-\epsilon'$ and $\mathbf{u}\cdot\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)\mid \mathbf{w}\cdot\mathbf{x}\in B_\phi]+\epsilon' \geq \mathbf{u}\cdot\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)\mid \mathbf{w}\cdot\mathbf{x}\in B_\phi]$.

Using the fact that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)\mid B]\,\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]$, we get

$$\mathbf{u}\cdot\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq \mathbf{u}\cdot\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\frac{\mathbf{Pr}_{\mathbf{x}\sim\widehat{\mathcal{D}}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]}{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]} - \epsilon'\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\widehat{\mathcal{D}}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]\,.$$

We need to show that $|\,\mathbf{Pr}_{\mathbf{x}\sim\widehat{\mathcal{D}}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi] - \mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]| \leq 2\,\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]$. To prove that, we use the following inequality, known as Dvoretzky–Kiefer–Wolfowitz (DKW) inequality.

Fact C.15 (DKW inequality (Naaman, 2021))**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$ and let $\widehat{\mathcal{D}}$ be its empirical with N samples. Denote $F_N(\mathbf{z}) = \mathbf{E}_{\mathbf{x}\sim\widehat{\mathcal{D}}_\mathbf{x}}[\prod_{i=1}^d \mathbb{1}\{\mathbf{x}_i\leq\mathbf{z}_i\}]$ and $F(\mathbf{z}) = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\prod_{i=1}^d \mathbb{1}\{\mathbf{x}_i\leq\mathbf{z}_i\}]$. Then, it holds that*

$$\mathbf{Pr}[\sup_{\mathbf{z}\in\mathbb{R}^d}(F_N(\mathbf{z})-F(\mathbf{z}))\geq\epsilon] \leq (N+1)de^{-2N\epsilon^2}\,.$$

Denote $q = \mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]$. Using the fact above for $N\geq O(\frac{\log(d/\delta)}{q^2})$, we have that $|\,\mathbf{Pr}_{\mathbf{x}\sim\widehat{\mathcal{D}}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi] - \mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]| \leq 2\,\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]$ with probability at least $1-\delta$. Therefore, with probability at least $1-2\delta$

$$\mathbf{u}\cdot\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq (1/2)\mathbf{u}\cdot\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] - 2\epsilon'\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi]\,.$$

Therefore, by setting $\epsilon'' = \epsilon'/(4q)$, we get that

$$\mathbf{u}\cdot\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \geq (1/2)\mathbf{u}\cdot\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] - \epsilon''\,.$$

Moreover, using a similar aproach as before, we bound the $\|\,\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2$. For any unit vector $\mathbf{u}$, we have that

$$\mathbf{u}\cdot\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] \leq 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{u}\cdot\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] + 2\epsilon'\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{w}\cdot\mathbf{x}\in B_\phi] \leq 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{u}\cdot\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)] + \epsilon''\,.$$

Using that $\|\,\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2 \leq \max_{\mathbf{u}\in\mathbb{R}^d,\|\mathbf{u}\|_2=1}|\mathbf{u}\cdot\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]|_2$, we get that $\|\,\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2 \leq 2\|\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w},\phi,\mathbf{x},y)]\|_2+\epsilon''$.

Note that in order to get one sample from the region B, you need $O(1/q)$ samples from $\mathcal{D}$, therefore overall you need $O\left(\frac{d\log(1/\delta)q}{\epsilon''^2}+\frac{\log(d/\delta)}{q^2}\right)$ samples from $\mathcal{D}$.