Learning a Single Neuron with Adversarial Label Noise via Gradient Descent

Ilias Diakonikolas ILIAS@CS.WISC.EDU
Vasilis Kontonis KONTONIS@WISC.EDU
Christos Tzamos TZAMOS@WISC.EDU
Nikos Zarifis ZARIFIS@WISC.EDU
UW Madison

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study the fundamental problem of learning a single neuron, i.e., a function of the form $\mathbf{x} \mapsto \sigma(\mathbf{w} \cdot \mathbf{x})$ for monotone activations $\sigma : \mathbb{R} \mapsto \mathbb{R}$, with respect to the L_2^2-loss in the presence of adversarial label noise. Specifically, we are given labeled examples from a distribution D on $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that there exists $\mathbf{w}^* \in \mathbb{R}^d$ achieving $F(\mathbf{w}^*) = \mathrm{opt}$, where $F(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. The goal of the learner is to output a hypothesis vector $\widetilde{\mathbf{w}}$ such that $F(\widetilde{\mathbf{w}}) = C\,\mathrm{opt} + \epsilon$ with high probability, where $C > 1$ is a universal constant. As our main contribution, we give efficient constant-factor approximate learners for a broad class of distributions (including log-concave distributions) and activation functions (including ReLUs and sigmoids). Concretely, for the class of isotropic log-concave distributions, we obtain the following important corollaries:

- For the logistic activation, i.e., $\sigma(t) = 1/(1 + e^{-t})$, we obtain the first polynomial-time constant factor approximation (even under the Gaussian distribution). Our algorithm has sample complexity $\widetilde{O}(d/\epsilon)$, which is tight within polylogarithmic factors.
- For the ReLU activation, i.e., $\sigma(t) = \max(0, t)$, we give an efficient algorithm with sample complexity $\widetilde{O}(d\,\mathrm{polylog}(1/\epsilon))$. Prior to our work, the best known constant-factor approximate learner had sample complexity $\widetilde{\Omega}(d/\epsilon)$.

In both of these settings, our algorithms are simple, performing gradient-descent on the (regularized) L_2^2-loss. The correctness of our algorithms relies on novel structural results that we establish, showing that (essentially all) stationary points of the underlying non-convex loss are approximately optimal.

Keywords: List of keywords

1. Introduction

1.1. Background and Motivation

The recent success of deep learning has served as a practical motivation for the development of provable efficient learning algorithms for various natural classes of neural networks. Despite extensive investigation, our theoretical understanding of the assumptions under which neural networks are provably efficiently learnable remains somewhat limited. Here we focus on arguably the simplest possible setting of learning a *single* neuron, i.e., a real-valued function of the form $\mathbf{x} \mapsto \sigma(\mathbf{w} \cdot \mathbf{x})$, where $\mathbf{w}$ is the weight vector of parameters and $\sigma : \mathbb{R} \to \mathbb{R}$ is a fixed non-linear and monotone activation function. Concretely, the learning problem is the following: Given i.i.d. samples from a distribution D on $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^d$ is the feature vector and $y \in \mathbb{R}$ is the corresponding

label, our goal is to learn the underlying function in L_2^2-loss. That is, the learner's objective is to output a hypothesis $h : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbf{E}_{(\mathbf{x},y) \quad D}[(h(\mathbf{x}) - y)^2]$ is as small as possible, compared to the minimum possible loss $\mathrm{opt} := \min_{\mathbf{w} 2 \mathbb{R}^d} \mathbf{E}_{(\mathbf{x},y) \quad D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. Settings of particular interest for the activation σ include the ReLU and sigmoid functions, corresponding to $\sigma(u) = \mathrm{ReLU}(u) := \max\{0, u\}$ and $\sigma(u) := 1/(1 + \exp(-u))$ respectively. Recall that a learning algorithm is called *proper* if the hypothesis h is restricted to be of the form $h_{\widehat{\mathbf{w}}}(\mathbf{x}) = \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x})$. Throughout this paper, we focus on developing efficient proper learners.

In the realizable case, i.e., when the labels y are consistent with a function in the class, the above learning problem is known to be solvable in polynomial time for a range of activation functions. A line of work, see, e.g., Kalai and Sastry (2009); Soltanolkotabi (2017); Yehudai and Shamir (2020) and references therein, has shown that simple algorithms like gradient-descent efficiently converge to an optimal solution under additional assumptions on the marginal distribution $D_{\mathbf{x}}$ on examples.

In this work, we focus on the *agnostic* learning model, where no realizability assumptions are made on the distribution D. Roughly speaking, the agnostic model corresponds to learning in the presence of adversarial label noise.

Definition 1 (Learning Single Neurons with Adversarial Noise) *Fix $\epsilon, W > 0$, $\delta \in (0, 1)$, and a class of distributions $\mathcal{G}$ on $\mathbb{R}^d$. Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an activation function and D a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that its $\mathbf{x}$-marginal belongs in $\mathcal{G}$. We define the population L_2^2-loss as $F^{D,\sigma}(\mathbf{w}) \quad , \quad (1/2)\, \mathbf{E}_{(\mathbf{x},y) \quad D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. We say that D is (ϵ, W)-corrupted if $\inf_{\|\mathbf{w}\|_2 \quad W} F^{D,\sigma}(\mathbf{w}) \leq \epsilon$. For some $C \geq 1$, a C-approximate learner is given ϵ, δ, W and i.i.d. labeled examples from D and outputs a function $h(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ such that, with probability at least $1 - \delta$, it holds $(1/2)\, \mathbf{E}_{(\mathbf{x},y) \quad D}[(h(\mathbf{x}) - y)^2] \leq C\epsilon$.*

Some comments are in order. First, we note that the parameter ϵ quantifies the degree of contamination — in the sense that the closest function in the class has L_2^2-loss ϵ. (Sometimes, this is denoted by opt in the relevant literature.) The parameter W quantifies the radius of the ball that contains an optimal vector $\mathbf{w}$. For uniformly bounded activations (e.g., sigmoids), one can remove the restriction on the norm of the weight vector $\mathbf{w}$, i.e., take $W = +\infty$. In this case, we call D ϵ-corrupted. When the distribution and activation are clear from the context, we will write $F(\mathbf{w})$ instead of $F^{D,\sigma}(\mathbf{w})$.

Related Prior Work In this paper, we focus on developing efficient *constant-factor* approximate proper learners for a range of activation functions. For this to be possible in polynomial time, one needs to make some qualitative assumptions on the underlying marginal distribution on examples. Indeed, it is known (Diakonikolas et al., 2022) that in the distribution-free setting no constant-factor approximation is possible (even for improper learning), under cryptographic assumptions, for a range of activations including ReLUs and sigmoids. On the other hand, even under Gaussian marginals, achieving error $\mathrm{opt} + \epsilon$ (corresponding to $C = 1$ in Definition 1) requires time $d^{F(1/\epsilon)}$, for some function F with $\lim_{u \, ! \, 1} F(u) = \infty$ (Goel et al.; Diakonikolas et al., 2020c; Goel et al., 2020; Diakonikolas et al., 2021b). These hardness results motivate the design of constant-factor approximate learners under "well-behaved" distributions.

On the algorithmic side, Goel et al. (2017) gave an algorithm with error $\mathrm{opt} + \epsilon$ and runtime $\mathrm{poly}(d) 2^{\mathrm{poly}(1/\epsilon)}$ that succeeds as long as the distribution on examples is supported on the unit sphere. Frei et al. (2020) study ReLU activations under structured distributions and show that gradient-descent on the L_2 loss converges to a weight vector with error $O(d\, \epsilon)$. The most relevant prior work is by Diakonikolas et al. (2020a) who gave a $\mathrm{poly}(d/\epsilon)$-time constant-factor approximate

proper learner for ReLU activations under isotropic log-concave distributions. Their algorithm makes essential use of the ReLU activation. In fact, Diakonikolas et al. (2020a) asked whether efficient constant-factor approximations exist for other activations, including sigmoids. In this work, we answer this open question in the affirmative.

The aforementioned discussion motivates the following broad question:

Is there an efficient constant-factor approximate learner
for single neurons under well-behaved distributions?

In this work, we answer this question in the affirmative for a range of activations including ReLUs and sigmoids and a variety of well-behaved distributions. In fact, we show that a simple gradient-based method on the L_2^2-loss suffices.

1.2. Our Results

Distributional Assumptions We develop algorithms that are able to learn single neurons under a large class of structured distributions. We make mild distributional assumptions requiring only concentration, anti-concentration, and anti-anti-concentration on the $\mathbf{x}$-marginal of the examples. In particular, we consider the following class of well-behaved distributions.

Definition 2 (Well-behaved Distributions) *Let $L, R > 0$. An isotropic (i.e., zero mean and identity covariance) distribution $D_{\mathbf{x}}$ on $\mathbb{R}^d$ is called (L, R)-well-behaved if for any projection $(D_{\mathbf{x}})_V$ of $D_{\mathbf{x}}$ onto a subspace V of dimension at most two, the corresponding pdf γ_V on $\mathbb{R}^2$ satisfies the following:*

- *For all $\mathbf{x} \in V$ such that $\|\mathbf{x}\|_1 \leq R$ it holds $\gamma_V(\mathbf{x}) \geq L$ (anti-anti-concentration).*
- *For all $\mathbf{x} \in V$ it holds that $\gamma_V(\mathbf{x}) \leq (1/L)(e^{Lk\mathbf{x}k_2})$ (anti-concentration and concentration).*

When the parameters L, R are bounded above by universal constants (independent of the dimension), we will simply say that the distribution $D_{\mathbf{x}}$ is well-behaved.

The class of well-behaved distributions is fairly broad. Specifically, isotropic log-concave distributions are well-behaved, i.e., they are (L, R)-well-behaved for some $L, R = O(1)$, see, e.g., Lovász and Vempala (2007); Klivans et al. (2009). Similar assumptions were introduced in Diakonikolas et al. (2020d) and have been used in various classification and regression settings (Diakonikolas et al., 2020f,e,b, 2021a; Frei et al., 2020, 2021; Zhang and Li, 2021; Zou et al., 2021).

Learning Sigmoidal Activations Our first main result holds for a natural class of activations that roughly have "sigmoidal" shape.

Definition 3 (Sigmoidal Activations) *Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be a non-decreasing activation function and $\tau, \mu, \xi > 0$. We say that σ is (τ, μ, ξ)-sigmoidal if it satisfies (a) $\sigma'(t) \geq \tau$, for all $t \in [-1, 1]$, and (b) $\sigma'(t) \leq \xi e^{-\mu|t|}$, for all $t \in \mathbb{R}$. We will simply say that an activation is "sigmoidal" when τ, μ, ξ are universal constants.*

Arguably the most popular sigmoidal activation is the logistic activation or sigmoid, corresponding to $\sigma(t) = 1/(1 + e^{-t})$. Other well-studied sigmoidal activations include the hyperbolic tangent, the Gaussian error function, and the ramp activation (see Figure 1). We note that all these activations satisfy the requirement of Definition 3 for some universal constants τ, μ, ξ. In what follows, we will simply refer to them as sigmoidal.

The most commonly used method to solve our learning problem in practice is to directly attempt to minimize the L_2^2-loss via (stochastic) gradient descent. Due to the non-convexity of the objective, this method is of a heuristic nature in general and comes with no theoretical guarantees, even in noiseless settings. In our setting, the situation is even more challenging due to the adversarial noise in the labels. Indeed, we show that the "vanilla" L_2^2-objective may contain bad local optima, even under Gaussian marginals. Specifically, even with an arbitrarily small amount of adversarial noise, the vanilla L_2^2 objective will have local-minima whose L_2^2-error is larger than $1/2$ (which is essentially trivial, since sigmoidal activations take values in $[-1, 1]$); see Proposition 9 and Figure 2.

Our main structural result for sigmoidal activations is that we can "correct" the optimization landscape of the L_2^2-loss by introducing a standard ℓ_2-regularization term. We prove the following theorem showing that any stationary point of the regularized L_2^2-loss is approximately optimal.

Theorem 4 (Informal: Landscape of Sigmoidals) *For sigmoidal activations and ϵ-corrupted well-behaved distributions, any (approximate) stationary point $\bar{\mathbf{w}}$ of the ℓ_2 regularized objective $F_\rho(\mathbf{w}) = F(\mathbf{w}) + (\rho/2)\|\mathbf{w}\|_2^2$ with $\rho = \Theta(\epsilon^3)$, satisfies $F(\bar{\mathbf{w}}) = O(\epsilon)$.*

Standard gradient methods (such as SGD) are known to efficiently converge to stationary points of non-convex objectives under certain assumptions. By running any such method on our regularized objective, we readily obtain an efficient algorithm that outputs a weight vector with L_2^2-loss $O(\epsilon)$. This is already a new result in this context. Yet, black-box application of optimization results for finding stationary points of non-convex functions would resulting in a sample complexity with suboptimal dependence on ϵ, up to polynomial factors.

Aiming towards an algorithmic result with near-optimal sample complexity, we perform a "white-box" analysis of gradient descent, leveraging the optimization landscape of sigmoidals. Specifically, we show that "vanilla" gradient descent, with a fixed step size, finds an approximately optimal solution when run on the empirical (regularized) L_2^2-loss with a near-optimal number of samples.

Theorem 5 (Informal: Learning Sigmoidals via Gradient Descent) *For sigmoidal activations and ϵ-corrupted well-behaved distributions, gradient descent on the empirical regularized loss $\widehat{F}_\rho(\cdot)$ with $N = \widetilde{\Theta}(d/\epsilon)$ samples, converges, in $\mathrm{poly}(1/\epsilon)$ iterations, to a vector $\widetilde{\mathbf{w}}$ satisfying $F(\widetilde{\mathbf{w}}) \leq O(\epsilon)$ with high probability.*

Theorem 5 gives the first efficient constant-factor approximate learner for sigmoid activations in the presence of adversarial label noise, answering an open problem of Diakonikolas et al. (2020a). As an additional bonus, our algorithm is simple and potentially practical (relying on gradient-descent), has near-optimal sample complexity (see Lemma 66), and succeeds for a broad family of bounded activation functions (i.e., the ones satisfying Definition 3).

For simplicity of the exposition, we have restricted our attention to sigmoidal activations (corresponding to τ, μ, ξ being universal constants in Definition 3) and well-behaved distributions (corresponding to $L, R = O(1)$ in Definition 2). In the general case, we note that the complexity of our algorithm is polynomial in the parameters L, R, τ, μ, ξ, see Theorem 22.

Learning Unbounded Activations We next turn our attention to activation functions that are not uniformly bounded. The most popular such activation is the ReLU function $\sigma(t) = \mathrm{Relu}(t) = \max(0, t)$. Our algorithmic results apply to the following class of unbounded activations.

Definition 6 (Unbounded Activations) *Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be a non-decreasing activation function and $\alpha, \lambda > 0$. We say that σ is (α, λ)-unbounded if it satisfies (a) σ is λ-Lipschitz and (b) $\sigma^0(t) \geq \alpha$, for all $t \in [0, +\infty)$. We will simply say that an activation is unbounded when the parameters α, λ are universal constants.*

We use the term "unbounded" for these activations, as they tend to ∞ as $t \to +\infty$. Most well-known unbounded activation functions such as the ReLU, Leaky-ReLU, ELU are (α, λ)-unbounded for some absolute constants $\alpha, \lambda > 0$. For example, the ReLU activation is $(1, 1)$-unbounded.

Our main structural result for unbounded activations is that all stationary points $\mathbf{w}$ of the L_2^2-loss that lie in the halfspace $\mathbf{w} \cdot \mathbf{w}^* \geq 0$, where $\mathbf{w}^*$ is the optimal weight vector, are approximately optimal. In more detail, we establish the following result.

Theorem 7 (Informal: Landscape of Unbounded Activations) *For unbounded activations and ϵ-corrupted well-behaved distributions, any stationary point $\bar{\mathbf{w}}$ of $F(\mathbf{w})$ with $\mathbf{w} \cdot \mathbf{w}^* \geq 0$ satisfies $F(\bar{\mathbf{w}}) = O(\epsilon)$, and any $\mathbf{w}$ with $\mathbf{w} \cdot \mathbf{w}^* \geq 0$ and $F(\mathbf{w}) = \Omega(\epsilon)$, satisfies $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) > 0$.*

We remark that the constant in the error guarantee of the above theorem only depends on the (universal) constants of the distribution and the activation and not on the radius W of Definition 1. Interestingly, Theorem 7 does not preclude the existence of suboptimal stationary points. On the other hand, similarly to the case of sigmoidal activations, our structural result can be readily combined with "black-box" optimization to efficiently find a constant-factor approximately optimal weight vector $\mathbf{w}$ (see Section 3). To obtain near-optimal sample complexity, we again perform an "white-box" analysis of (approximate) gradient descent, showing that simply by initializing at $\mathbf{0}$ we can avoid bad stationary points.

Theorem 8 (Informal: Learning Unbounded Activations via Gradient Descent) *For unbounded activations and (ϵ, W)-corrupted well-behaved distributions, (approximate) gradient descent on the L_2^2-loss $F(\cdot)$ with sample size $N = \widetilde{\Theta}(dW^2 \max(\mathrm{polylog}(W/\epsilon), 1))$ and $\mathrm{polylog}(1/\epsilon)$ iterations, converges to a vector $\widetilde{\mathbf{w}} \in \mathbb{R}^d$, satisfying $F(\widetilde{\mathbf{w}}) \leq O(\epsilon)$ with high probability.*

Theorem 8 is a broad generalization of the main result of Diakonikolas et al. (2020a), which gave a constant-factor approximation (using a different approach) for the special case of ReLU activations under isotropic log-concave distributions. While the result of Diakonikolas et al. (2020a) was tailored to the ReLU activation, Theorem 8 works for a fairly broad class of unbounded activations (and a more general class of distributions). A key conceptual difference between the two results lies in the overall approach: The prior work Diakonikolas et al. (2020a) used a *convex* surrogate for optimization. In contrast, we leverage our structural result and directly optimize the natural non-convex objective.

Additionally, Theorem 8 achieves significantly better (and near-optimal) sample complexity as a function of both d and $1/\epsilon$. In more detail, the algorithm of Diakonikolas et al. (2020a) had sample complexity $\Omega(d/\epsilon)$ (their results are phrased for the special case that $W = 1$), while our algorithm has sample complexity $\widetilde{O}(d)\mathrm{polylog}(1/\epsilon)$ — i.e., near-linear in d and polylogarithmic in $1/\epsilon$.

2. Preliminaries

Basic Notation For $n \in \mathbb{Z}_+$, let $[n] := \{1, \ldots, n\}$. We use small boldface characters for vectors and capital bold characters for matrices. For $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, $\mathbf{x}_i$ denotes the i-th coordinate of

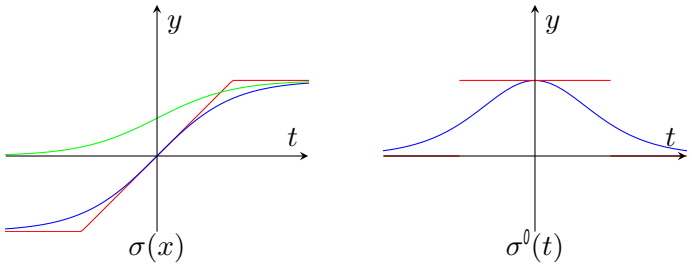


Figure 1: The $\tanh(\cdot)$ activation is $(0.4, 1, 1.4)$-sigmoidal. The ramp activation is $(1, 1, 3)$-sigmoidal. The logistic activation is $(0.19, 1, 1)$-sigmoidal. In the right figure, we plot the derivative of the ramp activation that is simply a rectangular function $\mathtt{1}\{|t| \leq 1\}$ (drawn in red) and the derivative of $\tanh(t)$ (drawn in blue). Notice that both decay at least exponentially fast: the derivative of the ramp activation is non-zero only in the interval $[-1, 1]$ and the derivative of $\tanh(t)$ decays exponentially fast, i.e., it is always smaller than $(4/3)e^{-|t|}$.

$\mathbf{x}$, 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 $\mathbf{x}, \mathbf{y}$. We slightly abuse notation and denote $\mathbf{e}_i$ the i-th standard basis vector in $\mathbb{R}^d$. We will use $\mathtt{1}_A$ to denote the characteristic function of the set A, i.e., $\mathtt{1}_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $\mathtt{1}_A(\mathbf{x}) = 0$ if $\mathbf{x} \notin A$. For vectors $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$, we denote $\mathbf{v}^{\perp \mathbf{u}}$ the projection of $\mathbf{v}$ into the subspace orthogonal to $\mathbf{u}$, furthermore, we denote $\mathbf{v}^{\parallel \mathbf{u}}$ the projection of $\mathbf{v}$ into the direction $\mathbf{u}$, i.e., $\mathbf{v}^{\parallel \mathbf{u}} := ((\mathbf{v} \cdot \mathbf{u})\mathbf{u})/\|\mathbf{u}\|_2^2$.

Asymptotic Notation We use the standard $O(\cdot), \Theta(\cdot), \Omega(\cdot)$ asymptotic notation. We also use $\widetilde{O}(\cdot)$ to omit poly-logarithmic factors. We write $E \,\&\, F$, two non-negative expressions E and F to denote that *there exists* some positive universal constant $c > 0$ (independent of the variables or parameters on which E and F depend) such that $E \geq c\,F$. In other words, $E = \Omega(F)$. For non-negative expressions E, F we write $E \gg F$ to denote that $E \geq C\,F$, where $C > 0$ is a *sufficiently large* universal constant (again independent of the parameters of E and F). The notations $\lesssim, \ll$ are defined similarly.

Probability Notation We use $\mathbf{E}_{x \sim D}[x]$ for the expectation of the random variable x according to the distribution 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. For $(\mathbf{x}, y)$ distributed according to D, we denote $D_{\mathbf{x}}$ to be the distribution of $\mathbf{x}$ and D_y to be the distribution of y. For unit vector $\mathbf{v} \in \mathbb{R}^d$, we denote $D_{\mathbf{v}}$ the distribution of $\mathbf{x}$ on the direction $\mathbf{v}$, i.e., the distribution of $\mathbf{x}_{\mathbf{v}}$. For a set B and a distribution D, we denote D_B to be the distribution D conditional on B.

3. The Landscape of the L_2^2 Loss

In this section, we present our results on the landscape of the L_2^2 loss for the sigmoidal activation functions of Definition 3 and the unbounded activation functions of Definition 6. Before we proceed, we remark again that Definition 3 models a general class of bounded activation functions, including, for example, the logistic activation, $\sigma(t) = 1/(1 + e^{-t})$, the hyperbolic tangent, $\sigma(t) = \tanh(t)$, the Gaussian error function, and the ramp activation, see Figure 1.

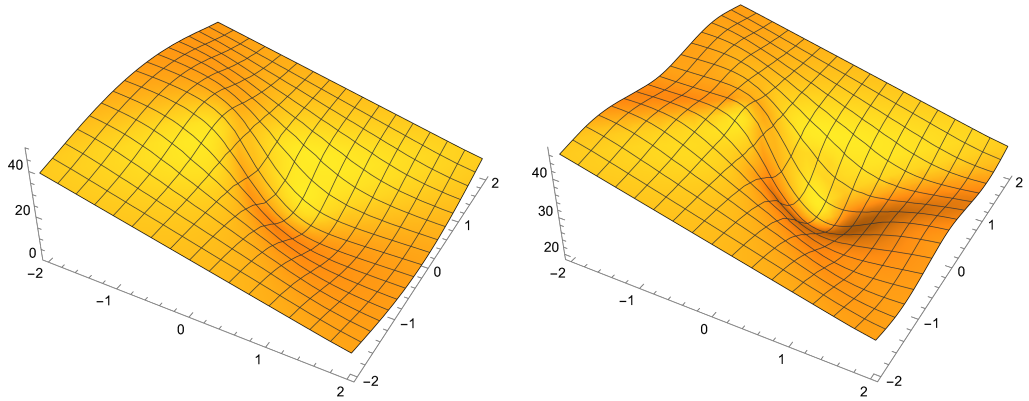


Figure 2: The non-convex optimization landscape of the "vanilla" L_2^2 loss. The activation is a ramp function $r(t)$, defined in Equation (2). The $\mathbf{x}$-marginal is the uniform distribution on the square $[-2,2] \times [-2,2]$. The "true" underlying weight vector is $\mathbf{w}^* = (1,0)$. In the left figure, we plot the noiseless population objective $F(\mathbf{w})$ as a function of $\mathbf{w} \in [-2,2] \times [-2,2]$, where $y = r(\mathbf{w}^* \cdot \mathbf{x})$. We observe that even though the objective is non-convex, in this case, the only stationary point is the true weight vector $\mathbf{w}^*$. In the right figure, we introduce noise and observe that the objective has another stationary point, which in fact is $-\mathbf{w}^*$. Finally, we notice that the landscape becomes more "flat" as we move away from the origin and the "bad" stationary point is a local-minimum that the "noise" was able to create in a particularly flat region.

We first show that we can construct noisy instances that have "bad" stationary points, i.e., local-minima whose L_2^2 loss is $\omega(\epsilon)$. Perhaps surprisingly, this is the case even when the underlying $\mathbf{x}$-marginal is the standard normal and the level of corruption ϵ of the corresponding labeled instance D is arbitrarily small. Moreover, the sigmoidal activation used in the construction of the counterexample is very simple (in particular, we use the ramp activation). We show that, even though the constructed instance is only ϵ-corrupted, there exists a local minimum whose L_2^2 loss is at least $\omega(\epsilon)$ (and in fact larger than some universal constant). The proof of the following result can be found in Appendix A; an example of the noisy L_2^2 landscape is shown in Figure 2.

Proposition 9 (Vanilla L_2^2 has "Bad" Local-Minima) *For any $\epsilon \in (0,1]$, there exists a well-behaved sigmoidal activation $\sigma(\cdot)$ and an ϵ-corrupted distribution D on $\mathbb{R}^d \times \{\pm 1\}$ with standard Gaussian $\mathbf{x}$-marginal such that $F^{D,\sigma}(\cdot)$ has a local minimum $\mathbf{u}$ with $F^{D,\sigma}(\mathbf{u}) \geq 1/2$.*

Our positive result shows that by using a regularized version of the L_2^2 loss, we can guarantee that all stationary points have error within a constant multiple of ϵ. Before stating our formal result, we first define the norm that we will use frequently together with its corresponding dual norm. In fact, we show that this norm characterizes the landscape of the (regularized) L_2^2 loss, in the sense that minimizing the gradient with respect to its dual norm will give a point with small error.

Definition 10 ($\mathbf{w}$-weighted Euclidean Norm) *Given some vector $\mathbf{u} \in \mathbb{R}^d$, we define its weighted Euclidean norm with respect to a non-zero vector $\mathbf{w} \in \mathbb{R}^d$ to be $\|\mathbf{u}\|_{\mathbf{w}} = \frac{\|\mathrm{proj}_{\mathbf{w}}\mathbf{u}\|_2}{\|\mathbf{w}\|_2^{3/2}} + \frac{\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{u}\|_2}{\|\mathbf{w}\|_2^{1/2}}$. We also define the dual norm of $\|\cdot\|_{\mathbf{w}}$ as follows: $\|\mathbf{v}\|_{*,\mathbf{w}} = \max(\|\mathrm{proj}_{\mathbf{w}}\mathbf{v}\|_2\|\mathbf{w}\|_2^{3/2}, \|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{v}\|_2\|\mathbf{w}\|_2^{1/2})$.*

The main intuition behind the norm of Definition 10 is the following: for well-behaved distributions and sigmoidal activations, the "noiseless" L_2^2 loss $\mathbf{E}_{\mathbf{x}\ D}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{w}\ \cdot\mathbf{x}))^2]$ behaves similarly to the non-convex function $\mathbf{w}\mapsto\|\mathbf{w}-\mathbf{w}\ \|_{\mathbf{w}}^2$. For more details and intuition on why this norm naturally appears in our results, we refer to (the proofs of) Lemma 17 and Lemma 43.

Remark 11 (Distribution/Activation parameters) Before we present our main structural result, we would like to revisit the parameters of Definition 2 and Definition 3. We observe that an (L,R)-well-behaved distribution is also $(L^{\prime},R^{\prime})$-well-behaved for any $L^{\prime}\leq L, R^{\prime}\leq R$. Therefore, without loss of generality (and to simplify the presentation), we shall assume that $L,R\in(0,1]$. For the same reason, for sigmoidal activations, we will assume that $\xi\in[1,\infty)$ and $\tau,\mu\in(0,1]$. Similarly, for unbounded activations we assume $\lambda\in[1,\infty),\alpha\in(0,1]$.

We now state our main structural result for sigmoidal activations, namely that all stationary points of the ℓ_2-regularized L_2^2 objective are approximately optimal solutions. Its proof can be found in Appendix A.

Theorem 12 (Stationary Points of Sigmoidal Activations) *Let D be an ϵ-corrupted, (L,R)-well-behaved distribution on $\mathbb{R}^d\times\mathbb{R}$ and σ be a (τ,μ,ξ)-sigmoidal activation. Set $\kappa=L^6R^6\mu^3\tau^4/\xi^2$ and $\rho=C\epsilon^3/\kappa^5$, where $C>0$ is a universal constant, and define the ℓ_2-regularized objective as $F_\rho(\mathbf{w})=(1/2)\,\mathbf{E}_{(\mathbf{x},y)\ D}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2]+(\rho/2)\|\mathbf{w}\|_2^2$. For some sufficiently small universal constant $c>0$, we have the following*

- *If $\|\mathbf{w}\|_2\leq 2/R$ and $\|\nabla F_\rho(\mathbf{w})\|_2\leq c\sqrt{\epsilon}$, then $F(\mathbf{w})=\epsilon\cdot\mathrm{poly}(1/\kappa)$.*
- *If $\|\mathbf{w}\|_2\geq 2/R$ and $\|\nabla F_\rho(\mathbf{w})\|\ _{,\mathbf{w}}\leq c\sqrt{\epsilon}$, then $F(\mathbf{w})=\epsilon\cdot\mathrm{poly}(1/\kappa)$.*

Theorem 12 shows that the L_2^2 objective behaves differently when $\mathbf{w}$ lies close vs far from the origin. The landscape becomes more "flat" as we move further away from the origin, and in order to achieve loss ϵ we have to make the dual of the weighted $\mathbf{w}$-norm of the gradient small; for example, the component orthogonal to $\mathbf{w}$ has to be smaller than $\sqrt{\epsilon}/\|\mathbf{w}\|_2^{1/2}$.

At a high level, we prove Theorem 12 in two main steps. First, we analyze the population L_2^2 loss without a regularizer, and show that all "bad" stationary points are in fact at distance $\Omega(1/\epsilon)$ from the origin. Since the non-regularized (vanilla) gradient field becomes very flat far from the origin, adding a small amount of noise potentially creates bad local minima (see also Figure 2). We show that by adding an appropriate ℓ_2 regularizer, we essentially introduce radial gradients pulling towards the origin whose contribution is large enough to remove bad stationary points, while ensuring that the "good" stationary points do not change by a lot.

Before proceeding to the proof of Theorem 12, we first highlight an intricacy of the landscape of the L_2^2 loss for sigmoidal activations: the L_2^2 loss may be "minimized" for some direction of infinite length. Consider, for example, the case where the activation is the logistic loss $\sigma(t)=1/(1+e^{\ t})$, the label y is equal to $\mathrm{sign}(\mathbf{w}\ \cdot\mathbf{x})$ for some unit vector $\mathbf{w}$, and the $\mathbf{x}$-marginal is the standard normal distribution. Then, we have that for any fixed vector $\mathbf{w}$ it holds that $\mathbf{E}_{(\mathbf{x},y)\ D}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2]>0$ and

$$\lim_{t\ !\ +1}\ \mathop{\mathbf{E}}_{(\mathbf{x},y)\ D}[(\sigma(t\ \mathbf{w}\ \cdot\mathbf{x})-y)^2]=0\,.$$

We prove the following lemma showing that even though the true "minimizer" may be a vector of infinite length, there exist almost optimal solutions inside a ball of radius $O(1/\epsilon)$.

Lemma 13 (Radius of Approximate Optimality of Sigmoidal Activations) *Let D be an (L, R)-well-behaved distribution in $\mathbb{R}^d$ and let $\sigma(t)$ be a (τ, μ, ξ)-sigmoidal activation function. There exists a vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 \leq 1/\epsilon$ and $F(\mathbf{v}) \leq (1 + O(\frac{\xi}{\mu L}))\epsilon$.*

To prove Theorem 12 we will show that for most vectors $\mathbf{w}$ the gradients of vectors that have large error, i.e., $\mathbf{w}$ with $F(\mathbf{w}) \geq \Omega(\epsilon)$, will have large contribution towards some optimal vector $\mathbf{w}^*$, i.e., $\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) > 0$. This is a more useful property than simply proving that the norm of the gradient is large, and we will later use it to obtain an efficient algorithm. However, as we already discussed, this is not always the case: we can only show that the gradient field "pulls" towards some optimal solution only inside a ball of radius roughly $1/\epsilon$ around the origin (see Cases 1,2 of Proposition 14). Outside of this ball, we show that the regularizer will take effect and pull us back into the ball where the gradient field helps to improve the guess; this corresponds to Case 3 of Proposition 14. In particular, we show that the projection of the gradient of the L_2^2 objective on the direction $\mathbf{w} - \mathbf{w}^*$ is proportional to the standard ℓ_2 distance of $\mathbf{w} - \mathbf{w}^*$, when $\mathbf{w}$ is close to the origin, and proportional to the $\mathbf{w}$-weighted Euclidean distance of Definition 10, when $\mathbf{w}$ is far from the origin.

Proposition 14 (Gradient of the Regularized L_2^2 Loss) *Let D be an (L, R)-well-behaved distribution and define $F_\rho(\mathbf{w}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2] + (1/2)\rho\|\mathbf{w}\|_2^2$, where σ is a (τ, μ, ξ)-sigmoidal activation and $\rho > 0$. Let $\epsilon \in (0, 1)$ and let $\mathbf{w}^* \in \mathbb{R}^d$ such that $F(\mathbf{w}^*) \leq \epsilon$ and $\|\mathbf{w}^*\|_2 \leq U/\epsilon$, for some $U \geq 0$. Furthermore, set $\kappa = L^6 R^6 \mu^3 \tau^4/\xi^2$ and $\rho = C\epsilon^3/\kappa^5$, where $C \geq 0$ is a sufficiently large universal constant. There exists a universal constant $c^\emptyset > 0$, such that for any $\mathbf{w} \in \mathbb{R}^d$, we have:*

1. *When $\|\mathbf{w}\|_2 \leq 2/R$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq \sqrt{\epsilon}/(c^\emptyset \kappa^5)$, then $\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c^\emptyset \sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_2$.*
2. *When $2/R \leq \|\mathbf{w}\|_2 \leq c^\emptyset \kappa/\epsilon$ and either $\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}} \geq \sqrt{\epsilon}(U/(c^\emptyset \kappa^5))$ or $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}^*\|_2$, then $\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c^\emptyset \sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$.*
3. *When $\|\mathbf{w}\|_2 \geq c^\emptyset \kappa/(2\epsilon)$, then $\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} \geq c^\emptyset \sqrt{\epsilon}\|\mathbf{w}\|_{\mathbf{w}}$.*

To keep the presentation simple, we shall sketch the proof of the following proposition showing that the gradient field of the "vanilla" L_2^2 loss points towards some optimal solution $\mathbf{w}^*$, as long as the guess $\|\mathbf{w}\|_2$ is not very far from the origin. We refer to Appendix A for the proof of Proposition 14.

Proposition 15 (Gradient of the "Vanilla" L_2^2 Loss (Inside a Ball)) *Let D be an (L, R)-well-behaved distribution and let σ be a (τ, μ, ξ)-sigmoidal activation. Let $\epsilon \in (0, 1)$ be smaller than a sufficiently small multiple of $L^2 R^6 \tau^4/\xi^2$ and let $\mathbf{w}^* \in \mathbb{R}^d$ with $F(\mathbf{w}^*) \leq \epsilon$. Set $\kappa = L^6 R^6 \mu^3 \tau^4/\xi^2$. There exists a universal constant $c^\emptyset > 0$, such that for any $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq c^\emptyset \kappa/\epsilon$, we have:*

1. *When $\|\mathbf{w}\|_2 \leq 2/R$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq \sqrt{\epsilon}\xi/(c^\emptyset L R^4 \tau^2)$, then $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c^\emptyset \sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_2$.*
2. *When $\|\mathbf{w}\|_2 \geq 2/R$ and either $\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}} \geq \sqrt{\epsilon/\kappa}$ or $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}^*\|_2$, then $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c^\emptyset \sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$.*

Proof Sketch of Proposition 15 Let $\mathbf{w}^* \in \mathbb{R}^d$ be a target weight vector with $F(\mathbf{w}^*) \leq \epsilon$. As we discussed, we want to show that when $\mathbf{w}$ is far from $\mathbf{w}^*$ (in $\mathbf{w}$-weighted Euclidean norm), then the inner product $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*)$ is strictly positive. We can decompose this inner product in two parts: I_1 that depends on how "noisy" the labels are and I_2 that corresponds to the contribution that we would have if all labels were "clean", i.e., $y = \sigma(\mathbf{w}^* \cdot \mathbf{x})$,

$$\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) = \underbrace{\mathbf{E}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})]}_{I_1} + \underbrace{\mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - \sigma(\mathbf{w}^* \cdot \mathbf{x}))\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})]}_{I_2} . \tag{1}$$

We crucially use the monotonicity of the activation $\sigma(\cdot)$: since $\sigma(\cdot)$ is non-decreasing, we have that $\sigma'(t) \geq 0$ and $(\sigma(\mathbf{w} \cdot \mathbf{x}) - \sigma(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x}) \geq 0$. These facts immediately imply that $I_2 \geq 0$. In what follows, we will show that I_2 is strictly positive.

In the worst case, the noisy term I_1 will be negative, i.e., the noise will try to make the gradient point in the wrong direction (move $\mathbf{w}$ "away" from the target $\mathbf{w}^*$). In the first part of the proof, we show that $|I_1|$ cannot be too large: the positive contribution I_2 is much larger than the contribution of the noisy term $|I_1|$. Apart from the monotonicity of the activation $\sigma(\cdot)$, we will also rely on the fact that if we view its derivative as a distribution, it satisfies anti-concentration and anti-anti-concentration properties, i.e., $\sigma'(t) \geq \tau$ for all $t \in [-1, 1]$ and $\sigma'(t) \leq \xi e^{-\mu j t j}$ for all $t \in \mathbb{R}$, see Definition 3. A simple sigmoidal activation is the ramp activation, defined as follows:

$$r(t) \;,\; (-1)\,\mathbf{1}\{t < -1\} + t\,\mathbf{1}\{|t| \leq 1\} + (+1)\,\mathbf{1}\{t > 1\} . \tag{2}$$

We observe that the ramp activation is $(1, e, 1)$-sigmoidal since its derivative is $r'(t) = 1\{|t| \leq 1\}$ and vanishes exactly outside the interval $[-1, 1]$. In the formal proof, we show how to reduce the analysis of general sigmoidal activations to the ramp activation. To keep this sketch simple, we will focus on the ramp activation.

Estimating the Contribution of the Noise We start by showing that the noise cannot affect the gradient by a lot, i.e., we bound the contribution of I_1. We prove the following lemma.

Lemma 16 *Let D be a well-behaved distribution. For any vector $\mathbf{w} \in \mathbb{R}^d$ it holds that $|I_1| \;.\; \sqrt{\epsilon}\,\min(\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}, \|\mathbf{w} - \mathbf{w}^*\|_2)$.*

Proof. (Sketch) Using the Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned} |I_1| &\leq \mathbf{E}[|(r(\mathbf{w}^* \cdot \mathbf{x}) - y)r'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})|] \\ &\leq (\mathbf{E}[(r(\mathbf{w}^* \cdot \mathbf{x}) - y)^2])^{1/2} (\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 r'(\mathbf{w} \cdot \mathbf{x})^2])^{1/2} \\ &= \sqrt{F(\mathbf{w}^*)}\,(\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 r'(\mathbf{w} \cdot \mathbf{x})^2])^{1/2} \\ &\leq \sqrt{\epsilon}\,(\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 r'(\mathbf{w} \cdot \mathbf{x})^2])^{1/2} . \end{aligned} \tag{3}$$

We proceed to bound the term $\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 r'(\mathbf{w} \cdot \mathbf{x})^2]$. Note that we can use the global upper bound on the derivative of the activation function, i.e., that $r'(t) \leq e$ for all $t \in \mathbb{R}$. However, this would only result in the bound $\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 r'(\mathbf{w} \cdot \mathbf{x})^2] \leq O(\|\mathbf{w} - \mathbf{w}^*\|_2^2)$. For sigmoidal activation functions, we need a tighter estimate that takes into account the fact that the functions

have exponential tails outside of the interval $[-1, 1]$. In particular, for the ramp function, we have that as $\|\mathbf{w}\|_2$ becomes larger the derivative $r'(\mathbf{w} \cdot \mathbf{x}) = \mathbf{1}\{|\mathbf{w} \cdot \mathbf{x}| \leq 1\}$ becomes a very thin band around the origin. Therefore, by the anti-concentration and anti-anti-concentration of well-behaved distributions (see Definition 2), the aforementioned integral decays to 0 as $\|\mathbf{w}\|_2 \to \infty$ at a rate of $1/\|\mathbf{w}\|_2$. Lemma 16 follows from combining Equation (3) along with the following lemma.

Lemma 17 *Let D be a well-behaved distribution. For any vectors $\mathbf{w}, \mathbf{w}^* \in \mathbb{R}^d$, it holds that* $\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 (r'(\mathbf{w} \cdot \mathbf{x}))^2] \gtrsim \min\left(\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}^2, \|\mathbf{w} - \mathbf{w}^*\|_2^2\right)$.

We will provide a proof of the above lemma because it highlights why the weighted Euclidean norm of Definition 10 captures the non-convex geometry of the optimization landscape.

Proof (Sketch) Here we assume that $D_{\mathbf{x}}$ is the standard d-dimensional Gaussian; see Lemma 35 for the formal version. First note that $\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 r'(\mathbf{w} \cdot \mathbf{x})^2] \leq \mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[((\mathbf{w} - \mathbf{w}^*) \cdot \mathbf{x})^2] = \|\mathbf{w} - \mathbf{w}^*\|_2^2$, where we used the fact that since the distribution $D_{\mathbf{x}}$ is isotropic, i.e., for any vector $\mathbf{u} \in \mathbb{R}^d$ it holds that $\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{u} \cdot \mathbf{x})^2] = \|\mathbf{u}\|_2^2$. This bound is tight when $\|\mathbf{w}\|_2$ is small (see Case 1 of Proposition 15). When $\|\mathbf{w}\|_2 \geq 1$, we see that the upper bound decays with $\|\mathbf{w}\|_2$. Let $\mathbf{q} = \mathbf{w} - \mathbf{w}^*$ and denote by G the one-dimensional standard Gaussian. We can decompose the difference $\mathbf{q}$ to its component parallel to $\mathbf{w}$: $\mathbf{q}^{\parallel \mathbf{w}}$ and its component orthogonal to $\mathbf{w}$: $\mathbf{q}^{\perp \mathbf{w}}$. From the Pythagorean theorem, we have that $(\mathbf{q} \cdot \mathbf{x})^2 = (\mathbf{q}^{\parallel \mathbf{w}} \cdot \mathbf{x})^2 + (\mathbf{q}^{\perp \mathbf{w}} \cdot \mathbf{x})^2$. Using this we get:

$$\mathbf{E}[(\mathbf{q} \cdot \mathbf{x})^2 \mathbf{1}\{|\mathbf{w} \cdot \mathbf{x}| \leq 1\}] = \mathbf{E}[(\mathbf{q}^{\parallel \mathbf{w}} \cdot \mathbf{x})^2 \mathbf{1}\{|\mathbf{w} \cdot \mathbf{x}| \leq 1\}] + \mathbf{E}[(\mathbf{q}^{\perp \mathbf{w}} \cdot \mathbf{x})^2) \mathbf{1}\{|\mathbf{w} \cdot \mathbf{x}| \leq 1\}]$$

$$= \|\mathbf{q}^{\parallel \mathbf{w}}\|_2^2 \, \mathbf{E}_{z \sim G}[z^2 \mathbf{1}\{|z| \leq \|\mathbf{w}\|_2^{-1}\}] + \|\mathbf{q}^{\perp \mathbf{w}}\|_2^2 \, \mathbf{Pr}_{z \sim G}[|z| \leq \|\mathbf{w}\|_2^{-1}] \gtrsim \frac{\|\mathbf{q}^{\parallel \mathbf{w}}\|_2^2}{\|\mathbf{w}\|_2^3} + \frac{\|\mathbf{q}^{\perp \mathbf{w}}\|_2^2}{\|\mathbf{w}\|_2} \gtrsim \|\mathbf{q}\|_{\mathbf{w}}^2 ,$$

where in the second equality we used that under the Gaussian distribution any two orthogonal directions are independent, and in the last inequality we used that $a^2 + b^2 \leq (a + b)^2$ for $a, b \geq 0$. Observe that the orthogonal direction is only scaled by roughly the probability of the slice, which is $1/\|\mathbf{w}\|_2$, while the parallel component (which is restricted in the interval $|z| \leq 1/\|\mathbf{w}\|_2$) is scaled by $1/\|\mathbf{w}\|_2^3$ (similarly to the fact that $\int_a^a t^2 dt = O(a^3)$). ■

Estimating the Contribution of the "Noiseless" Gradient We now bound from below the contribution of the "noiseless" gradient, i.e., the term I_2 of Equation (1). To bound I_2 from below, we consider several different cases depending on how far the vector $\mathbf{w}$ is from the target vector $\mathbf{w}^*$; the full proof is rather technical. In this sketch, we shall only consider the case where the angle between the weight vectors $\mathbf{w}$ and $\mathbf{w}^*$ is in $(0, \pi/2)$ and also assume that the norm of the guess $\mathbf{w}$ is larger than $2/R$, since this is the regime where the norm of the gradient behaves similarly to the weighted Euclidean norm of Definition 10. Notice that when $\mathbf{w}$ is close to the target $\mathbf{w}^*$, then its projection on the orthogonal complement of $\mathbf{w}$, i.e., $\|\mathrm{proj}_{\mathbf{w}^\perp} \mathbf{w}^*\|_2$, will be small and also its projection on the direction of $\mathbf{w}$, i.e., $\|\mathrm{proj}_{\mathbf{w}} \mathbf{v}\|_2$, will be close to $\|\mathbf{w}\|_2$. In this sketch, we will show how to handle the case where $\theta(\mathbf{w}, \mathbf{w}^*) \in (0, \pi/2)$, $\|\mathrm{proj}_{\mathbf{w}^\perp} \mathbf{w}^*\|_2 \leq 2/R$ and $\|\mathrm{proj}_{\mathbf{w}} \mathbf{w}^*\|_2 \leq 2\|\mathbf{w}\|_2$, i.e., the case where $\mathbf{w}$ and $\mathbf{w}^*$ are not extremely far apart.

Lemma 18 *Let D be a well-behaved distribution. For any vector $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 > 2/R$, $\|\mathrm{proj}_{\mathbf{w}^\perp} \mathbf{w}^*\|_2 \leq 2/R$, $\|\mathrm{proj}_{\mathbf{w}} \mathbf{w}^*\|_2 \leq 2\|\mathbf{w}\|_2$, $\theta(\mathbf{w}, \mathbf{w}^*) \in (0, \pi/2)$, it holds that $I_2 \gtrsim \|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}^2$.*

We can now finish the proof of (one case of) Proposition 15. From Lemma 16, if $\|\mathbf{w}\|_2 \geq 2/R$, it holds that the noisy gradient term $|I_1| \; . \; \sqrt{\epsilon}\|\mathbf{w} \; - \mathbf{w}\|_{\mathbf{w}}$. Using Lemma 18, there exists a universal constant $c > 0$, such that $I_1 + I_2 \geq c\|\mathbf{w} \; - \mathbf{w}\|_{\mathbf{w}}(\|\mathbf{w} \; - \mathbf{w}\|_{\mathbf{w}} - \sqrt{\epsilon}/c) \; \& \; \sqrt{\epsilon}\|\mathbf{w} \; - \mathbf{w}\|_{\mathbf{w}} \; ,$ where in the last inequality we used that $\|\mathbf{w} \; - \mathbf{w}\|_{\mathbf{w}} \gg \sqrt{\epsilon}$. This completes the proof sketch of Proposition 15. For the full proof, we refer to Appendix A.

4. The Landscape of the L_2^2-Loss for Unbounded Activations

For unbounded activations we essentially characterize the optimization landscape of the L_2^2-loss as a function of the weight vector $\mathbf{w}$. Specifically, our main structural theorem in this section establishes (roughly) the following: even though the population L_2^2-loss $F(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y) \; D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$ is not convex, any approximate stationary point $\mathbf{w}$ of $F(\mathbf{w})$ will have error close to that of the optimal weight vector $\mathbf{w} \; ,$ i.e., $F(\mathbf{w}) \leq O(\epsilon)$. In more detail, while there exist "bad" stationary points in this case, they all lie in a cone around $-\mathbf{w} \; ,$ i.e., in the opposite direction of the optimal weight vector. As we will see in Section 5, gradient descent initialized at the origin will always avoid such stationary points. Our main structural result for unbounded activations is as follows:

Theorem 19 (Stationary Points of (α, λ)-Unbounded Activations) *Let D be an (ϵ, W)-corrupted, (L, R)-well-behaved distribution in $\mathbb{R}^d$. Let σ be an (α, λ)-unbounded activation and let $F(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y) \; D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. Then, if for some $\mathbf{w} \in \mathbb{R}^d$, with $\mathbf{w} \cdot \mathbf{w} \; \geq 0$ and $\|\mathbf{w}\|_2 \leq W$ it holds $\|\nabla F(\mathbf{w})\|_2 \leq 2\lambda\sqrt{\epsilon}$, then $F(\mathbf{w}) \leq C\big(\frac{\lambda}{\alpha}\big)^4 \frac{1}{L^2R^8}\epsilon$, for some absolute constant $C > 0$.*

The proof of the theorem above can be found in Appendix B.

Remark 20 An important feature of the above theorem statement is that the error guarantee of the stationary points of the L_2^2-loss $F(\mathbf{w})$ does not depend on the size of the weight vector $\mathbf{w} \; .$ It only depends on the ratio of the constants λ and α (that for all activation functions discussed above is some small absolute constant). For the special case of ReLU activations, it holds that $\lambda/\alpha = 1$.

For the proof of Theorem 19, we need the following structural result for the gradient field of the L_2^2 loss for unbounded activations. We show that as long as $\|\mathbf{w} - \mathbf{w} \; \|_2$ is larger than roughly $\sqrt{\epsilon}$ and $\mathbf{w} \cdot \mathbf{w} \; \geq 0$, then the gradient at $\mathbf{w}$ has large projection in the direction of $\mathbf{w} - \mathbf{w} \; ,$ i.e., it "points in the right direction".

Proposition 21 *Let D be an (ϵ, W)-corrupted, (L, R)-well-behaved distribution and σ be an (α, λ)-unbounded activation. For any $\mathbf{w} \in \mathbb{R}^d$ with $\mathbf{w} \cdot \mathbf{w} \; \geq 0$ and $\|\mathbf{w} - \mathbf{w} \; \|_2 \geq C\lambda/(\alpha^2 LR^4)\sqrt{\epsilon}$, it holds $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w} \;) \; \& \; \alpha^2 LR^4 \|\mathbf{w} - \mathbf{w} \; \|_2^2$.*

We remark that one can readily use the above proposition to obtain an efficient algorithm via black-box optimization methods. To achieve this, we can simply start by finding a stationary point $\mathbf{w}^{(0)}$, and then repeat our search constraining on the set $\{\mathbf{w} : \mathbf{w} \cdot \mathbf{w}^{(0)} \leq 0\}$ to obtain some stationary point $\mathbf{w}^{(1)}$. Then we continue in the set $\{\mathbf{w}^{(0)} \cdot \mathbf{w} \leq 0, \mathbf{w}^{(1)} \cdot \mathbf{w} \leq 0\}$. Since for all the boundaries we have (using Proposition 21) that the gradient has non-zero projection onto the direction $\mathbf{w} - \mathbf{w} \; ,$ we obtain that by adding these linear constraints, we do not introduce new stationary points. It is not hard to see that after $O(d)$ such iterations, we obtain a stationary point that lies in the halfspace $\mathbf{w} \; \cdot \mathbf{w} \geq 0$, and therefore is an approximately optimal solution.

5. Optimizing the Empirical L_2^2-Loss

In this section, we show that given sample access to a distribution over labeled examples D on $\mathbb{R}^d \times \mathbb{R}$, we can optimize the L_2^2 loss via (approximate) gradient descent. This allows us to efficiently find a weight vector that achieves error $O(\epsilon)$. Ideally, we would like to perform gradient descent with the population gradients, i.e., $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta(\nabla F(\mathbf{w}^{(t)}))$. We do not have direct access to these gradients, but we show that it is not hard to estimate them using samples from D; see Algorithm 1. We present our main algorithmic result for learning the sigmoidal activations discussed in Section 3. The proof can be found in Appendix C.

Theorem 22 (Learning Sigmoidal Activations) *Let D be an ϵ-corrupted, (L,R)-well-behaved distribution on $\mathbb{R}^d \times \mathbb{R}$ and $\sigma(\cdot)$ be a (τ,μ,ξ)-sigmoidal activation. Set $\kappa = L^6R^6\mu^3\tau^4/\xi^2$ and let $c > 0$ be a sufficiently small absolute constant. Then gradient descent (Algorithm 1) with step size $\eta = c\epsilon^{2.5}$, regularization $\rho = (1/c)\epsilon^3/k^5$, truncation threshold $M = \xi/\mu$, $N = \widetilde{\Theta}(d/\epsilon \log(1/\delta))\ \mathrm{poly}(1/\kappa)$ samples, and $T = \mathrm{poly}(1/(\epsilon\kappa))$ iterations converges to a vector $\mathbf{w}^{(T)} \in \mathbb{R}^d$ that, with probability $1-\delta$, satisfies $F(\mathbf{w}^{(T)}) \leq \mathrm{poly}(1/\kappa)\ \epsilon$.*

Input: Iterations: T, N samples $(\mathbf{x}^{(i)}, y^{(i)})$ from D, step size: η, bound M, regularization ρ.
Output: A weight vector $\mathbf{w}^{(T)}$.

1. Let $\widehat{F}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(\sigma(\mathbf{w}\cdot\mathbf{x}^{(i)}) - \tilde{y}^{(i)})^2$, where $\widehat{y}^{(i)} = \mathrm{sign}(y^{(i)})\min(|y^{(i)}|, M)$
2. $\mathbf{w}^{(0)} \leftarrow \mathbf{0}$.
3. For $t = 0, \ldots, T$ do
   (a) $\mathbf{g}^{(t)} \leftarrow \nabla\widehat{F}(\mathbf{w}^{(t)})$.
   (b) $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta(\mathbf{g}^{(t)} + \rho\mathbf{w}^{(t)})$.

Algorithm 1: *Gradient Descent Algorithm for Optimizing the L_2^2-Loss for Sigmoidal Activations*

We now turn our attention to the case of unbounded activations. Perhaps surprisingly, we show that essentially the same algorithm (the only difference is that in this case we set the regularization parameter to 0) achieves sample complexity polylogarithmic in $1/\epsilon$. The proof of the following theorem can be found in Appendix C.

Theorem 23 (Unbounded Activations) *Let D be an (ϵ, W)-corrupted, (L,R)-well-behaved distribution on $\mathbb{R}^d \times \mathbb{R}$ and $\sigma(\cdot)$ be an (α,λ)-unbounded activation. Set $\kappa = \mathrm{poly}(LR\alpha/\lambda)/(W^2\log(W))$. The gradient descent Algorithm 2 with step size $\eta = \kappa\epsilon$, truncation threshold $M = \widetilde{O}((W/L)\ \max(\log(\lambda^2W^2/\epsilon), 1))$, $N = \widetilde{\Theta}((d/\kappa)\log(1/\delta)\max(\mathrm{poly}\log(1/\epsilon), 1))$ samples, and $T = \mathrm{poly}(\log(1/\epsilon), 1/\kappa)$ iterations converges to a vector $\mathbf{w}^{(T)} \in \mathbb{R}^d$ that, with probability $1-\delta$, satisfies $F(\mathbf{w}^{(T)})$. $\frac{1}{LR^4}\left(\frac{\lambda}{\alpha}\right)^4\epsilon$.*

Input: Iterations: T, sample access from D, step size: η, bound M.
Output: A weight vector $\mathbf{w}^{(T)}$.

1. $\mathbf{w}^{(0)} \leftarrow \mathbf{0}$.
2. For $t = 0, \ldots, T$ do
   (a) Use fresh N samples from D truncated at M to obtain $\mathbf{g}^{(t)}$, an approximation of the population gradient $\nabla F(\mathbf{w}^{(t)})$ (see Claim 55).
   (b) $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \mathbf{g}^{(t)}$.

Algorithm 2:*Gradient Descent Algorithm for Optimizing the L_2^2-Loss for Unbounded Activations*

Acknowledgments

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

I. Diakonikolas, S. Goel, S. Karmalkar, A. R. Klivans, and M. Soltanolkotabi. Approximation schemes for ReLU regression. In *Conference on Learning Theory, COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 1452–1485. PMLR, 2020a.

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

I. Diakonikolas, D. M. Kane, and N. Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under Gaussian marginals. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020c.

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

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

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

I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Efficiently learning halfspaces with tsybakov noise. *STOC*, 2021a.

I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the sq model. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021b.

I. Diakonikolas, D. Kane, P. Manurangsi, and L. Ren. Hardness of learning a single neuron with adversarial label noise. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

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

S. Frei, Y. Cao, and Q. Gu. Agnostic learning of halfspaces with gradient descent via soft margins. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

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

S. Goel, V. Kanade, A. R. Klivans, and J. Thaler. Reliably learning the relu in polynomial time. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 1004–1042, 2017.

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

A. T. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*. Citeseer, 2009.

A. R. Klivans, P. M. Long, and A. K. Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *13th International Workshop, RANDOM 2009*, pages 588–600, 2009.

P. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.

M. Soltanolkotabi. Learning ReLUs via gradient descent. In *Advances in neural information processing systems*, pages 2007–2017, 2017.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

G. Yehudai and O. Shamir. Learning a single neuron with gradient methods. In *Conference on Learning Theory, COLT*, 2020.

C. Zhang and Y. Li. Improved algorithms for efficient active learning halfspaces with massart and tsybakov noise. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021.

D. Zou, S. Frei, and Q. Gu. Provable robustness of adversarial training for learning halfspaces with noise. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

Appendix A. Omitted Proofs of Section 3

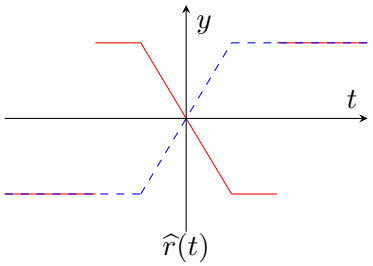
A.1. Proof of Proposition 9

We restate and prove the following proposition.

Proposition 24 (Bounded Activations Have "Bad" Stationary Points) *Fix $\epsilon \in (0,1]$ and let $r(t)$ be the ramp activation. There exists a distribution D on $\mathbb{R}^d \times \{\pm 1\}$ with standard Gaussian $\mathbf{x}$-marginal such that there exists a vector $\mathbf{v}$ with $F(\mathbf{v}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y)\ \ D}[(y - r(\mathbf{v}\cdot\mathbf{x}))^2] \le O(\epsilon)$ and a vector $\mathbf{u}$ with $\nabla F(\mathbf{u}) = 0$, $\nabla^2 F(\mathbf{u}) \preceq 0$, i.e., $\mathbf{u}$ is a local-minimum of F, and $F(\mathbf{u}) \ge 1/2$.*

Proof We define the following deterministic noise function

$$\widehat{r}(t) = \begin{cases} -1 & \text{if } t \le -2 \\ +1 & \text{if } -2 \le t \le -1 \\ -t & \text{if } -1 \le t \le 1 \\ -1 & \text{if } -1 \le t \le 2 \\ +1 & \text{if } t \ge 2 \end{cases}$$



In what follows, we denote by $\gamma(t) = \frac{1}{2\pi} e^{\ \ t^2/2}$ the density of the standard normal distribution. It suffices to consider noise that only depends on a single direction, i.e., $y(\mathbf{x}) = \widehat{r}(\mathbf{x}_1/\epsilon)$. We first show that there exists a vector with small L_2^2 loss. Take $\mathbf{v} = \mathbf{e}_1/\epsilon$, where $\mathbf{e}_1 = (1,0,\ldots,0)$ is the first vector of the standard orthogonal basis of $\mathbb{R}^d$. It holds that

$$\begin{aligned} F(\mathbf{v}) &= (1/2)\,\mathop{\mathbf{E}}_{\mathbf{x}\ \ D_{\mathbf{x}}}[(r(\mathbf{x}_1/\epsilon) - y(\mathbf{x}_1))^2] = (1/2)\,\mathop{\mathbf{E}}_{\mathbf{x}\ \ D_{\mathbf{x}}}[(r(\mathbf{x}_1/\epsilon) - \widehat{r}(\mathbf{x}_1/\epsilon))^2] \\ &= (1/2)\int_{\epsilon}^{\epsilon} (t/\epsilon + t/\epsilon)^2 \gamma(t)dt + \int_{2\epsilon}^{\epsilon} 4\gamma(t)dt \le \frac{\sqrt{2}}{\sqrt{\pi}\epsilon^2}\int_{\epsilon}^{\epsilon} t^2 dt + 4\epsilon = O(\epsilon)\,. \end{aligned}$$

We next show that there exists a "bad" stationary point $\mathbf{u}$, i.e., a $\mathbf{u}$ with $\nabla F(\mathbf{u}) = 0$ and $F(\mathbf{u}) \ge 1/2$. We have

$$\nabla F(\mathbf{u}) = \mathbf{E}[(r(\mathbf{u}\cdot\mathbf{x}) - y(\mathbf{x}))r'(\mathbf{u}\cdot\mathbf{x})\mathbf{x}]\,.$$

We shall take $\mathbf{u} = -\mathbf{e}_1/\epsilon$, i.e., the opposite direction of the almost optimal vector $\mathbf{v}$ that we used above. Using the coordinate-wise independence of the Gaussian distribution, we have that for every orthogonal direction $\mathbf{e}_i$, $i \ge 2$, it holds that $\nabla F(\mathbf{u})\cdot\mathbf{e}_i = 0$. For the direction $\mathbf{e}_1$, we obtain

$$\begin{aligned} \nabla F(\mathbf{u})\cdot\mathbf{e}_1 &= \mathbf{E}[(r(-\mathbf{x}_1/\epsilon) - \widehat{r}(\mathbf{x}_1/\epsilon))\,\mathbf{1}\{|\mathbf{x}_1| \le \epsilon\}\mathbf{x}_1] \\ &= \mathbf{E}[(r(-\mathbf{x}_1/\epsilon) - \widehat{r}(-\mathbf{x}_1/\epsilon))\,\mathbf{1}\{|\mathbf{x}_1| \le \epsilon\}\mathbf{x}_1] = 0\,. \end{aligned}$$

We next proceed to show that $\mathbf{u}$ is a local minimum of the population L_2^2-loss F. We compute the Hessian of F at $\mathbf{u}$. In what follows, we denote by $\delta(t)$ the standard Dirac delta function. We have that

$$\nabla^2 F(\mathbf{u}) = \mathop{\mathbf{E}}_{\mathbf{x}\ \ D_{\mathbf{x}}}[(r'(\mathbf{u}\cdot\mathbf{x}))^2\mathbf{x}\mathbf{x}^T] + \mathop{\mathbf{E}}_{\mathbf{x}\ \ D_{\mathbf{x}}}[(r(\mathbf{u}\cdot\mathbf{x}) - y(\mathbf{x}))(\delta(\mathbf{u}\cdot\mathbf{x}+1) - \delta(\mathbf{u}\cdot\mathbf{x}-1))\mathbf{x}\mathbf{x}^T]\,.$$

For $i \neq j$, using the fact that the Gaussian marginals are independent, we have that $(\nabla^2 F(\mathbf{u}))_{ij} = 0$. We next observe that the second term of the Hessian vanishes. We have

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(r(\mathbf{u}\cdot\mathbf{x}) - y(\mathbf{x}))(\delta(\mathbf{u}\cdot\mathbf{x}+1) - \delta(\mathbf{u}\cdot\mathbf{x}-1))\mathbf{x}_i^2] \\ &\qquad = \mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(r(-\mathbf{x}_1/\epsilon) - r(\mathbf{x}_1/\epsilon))(\delta(\mathbf{x}_1/\epsilon+1) - \delta(\mathbf{x}_1/\epsilon-1))\mathbf{x}_i^2] = 0\,,\end{aligned}$$

where we used the fact that both $r(-\mathbf{x}_1/\epsilon), \widehat{r}(\mathbf{x}_1/\epsilon)$ are continuous at $\mathbf{x}_1 = \pm\epsilon$, and it holds $r(\pm 1) = \widehat{r}(\pm 1)$.

To complete the proof, we need to show that the L_2^2 loss of $\mathbf{u}$ is large. It holds that

$$\begin{aligned}F(\mathbf{u}) &= \frac{1}{2}\,\mathbf{E}[(r(-\mathbf{x}_1/\epsilon) - r^{\emptyset}(\mathbf{x}_1/\epsilon))^2] \\ &\geq \frac{1}{2}\left(\int_{1}^{\epsilon}(r(-t/\epsilon) - r^{\emptyset}(t/\epsilon))^2\gamma(t)dt + \int_{\epsilon}^{+1}(r(-t/\epsilon) - r^{\emptyset}(t/\epsilon))^2\gamma(t)dt\right) \\ &\geq \frac{1}{2}\left(\int_{1}^{\epsilon} 4\gamma(t)dt + \int_{\epsilon}^{+1} 4\gamma(t)dt\right) \geq 4\int_{1}^{1}\gamma(t)dt \geq 1/2\,.\end{aligned}$$

This completes the proof of Proposition 9. ■

Proof of Lemma 13

We restate and prove the following lemma.

Lemma 25 (Radius of Approximate Optimality of Sigmoidal Activations) *Let D be an (L, R)-well-behaved distribution in $\mathbb{R}^d$ and let $\sigma(t)$ be a (τ, μ, ξ)-sigmoidal activation function. There exists a vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 \leq 1/\epsilon$ and $F(\mathbf{v}) \leq (1 + O(\frac{\xi}{\mu L}))\epsilon$.*

Proof We first observe that since $F(\mathbf{w})$ is a continuous function of $\mathbf{w}$, there exists a vector $\mathbf{w}$ with $\|\mathbf{w}\|_2 < +\infty$ such that $F(\mathbf{w}) \leq 2\epsilon$. If $\epsilon \geq 1$, then the zero vector achieves 4ϵ error since, using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have that $F(\mathbf{0}) = \mathbf{E}[y^2] \leq 2\,\mathbf{E}[(y - \sigma(\mathbf{w}\cdot\mathbf{x}))^2] + 2\,\mathbf{E}[\sigma(\mathbf{w}\cdot\mathbf{x})^2] \leq 2\epsilon + 2\epsilon \leq 4\epsilon$. Denote $1/\epsilon := \rho$. If $\|\mathbf{w}\|_2 \leq \rho$, then we are done. If $\|\mathbf{w}\|_2 > \rho$, then we consider a scaled down version of $\mathbf{w}$, namely, the vector $\mathbf{v} = \rho\,\mathbf{w}/\|\mathbf{w}\|_2$. Using again the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$F(\mathbf{v}) \leq 2F(\mathbf{w}) + 2\mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\sigma(\mathbf{v}\cdot\mathbf{x}) - \sigma(\mathbf{w}\cdot\mathbf{x}))^2]\,.$$

Since $\mathbf{v}$ is parallel to $\mathbf{w}$ and its norm is smaller than $\|\mathbf{w}\|$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\sigma(\mathbf{v}\cdot\mathbf{x}) - \sigma(\mathbf{w}\cdot\mathbf{x}))^2] &\leq \frac{1}{L}\int_{1}^{+1}(\sigma(\|\mathbf{v}\|_2 t) - \sigma(\|\mathbf{w}\|_2|t|))^2 dt \\ &\leq \frac{1}{L}\int_{1}^{+1}\Big(\int_{k\mathbf{v}kt}^{+1}\sigma^{\emptyset}(z)dz\Big)^2 dt \\ &\leq \frac{1}{L}\int_{1}^{+1}(\xi e^{\ \mu k\mathbf{v}kjtj})^2 dt \\ &= \frac{\xi}{\mu L\|\mathbf{v}\|}\,,\end{aligned}$$

where we used the fact that the density of 1-dimensional marginals of $D_{\mathbf{x}}$ is bounded from above by $1/L$, see Definition 2, and the fact that $\sigma^{\prime}(t) \leq \xi e^{-\mu|t|}$. We see that by choosing $\|\mathbf{v}\|_2 = 1/\epsilon$ it holds that $\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\sigma(\mathbf{v} \cdot \mathbf{x}) - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] = O(\xi/(\mu L))\epsilon$. Thus, combining the above we have that there always exists a $\mathbf{v}$ with $\|\mathbf{v}\|_2 = O(1/\epsilon$ that achieves L_2^2 loss $F(\mathbf{v}) \leq (1 + O(\frac{\xi}{\mu L}))\epsilon$. ■

A.2. Proof of Proposition 14

We restate and prove the following proposition.

Proposition 26 (Gradient of the Regularized L_2^2 Loss) *Let D be an (L, R)-well-behaved distribution and define $F_\rho(\mathbf{w}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y) \sim D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2] + (1/2)\rho\|\mathbf{w}\|_2^2$, where σ is a (τ, μ, ξ)-sigmoidal activation and $\rho > 0$. Let $\epsilon \in (0, 1)$ and let $\mathbf{w}^* \in \mathbb{R}^d$ such that $F(\mathbf{w}^*) \leq \epsilon$ and $\|\mathbf{w}^*\|_2 \leq U/\epsilon$, for some $U \geq 0$. Furthermore, set $\kappa = L^6 R^6 \mu^3 \tau^4/\xi^2$ and $\rho = C\epsilon^3/\kappa^5$, where $C \geq 0$ is a sufficiently large universal constant. There exists a universal constant $c^{\prime} > 0$, such that for any $\mathbf{w} \in \mathbb{R}^d$, we have:*

1. *When $\|\mathbf{w}\|_2 \leq 2/R$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq \sqrt{\epsilon}/(c^{\prime}\kappa^5)$, then $\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c^{\prime}\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_2$.*
2. *When $2/R \leq \|\mathbf{w}\|_2 \leq c^{\prime}\kappa/\epsilon$ and either $\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}} \geq \sqrt{\epsilon}(U/(c^{\prime}\kappa^5))$ or $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}^*\|_2$, then $\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c^{\prime}\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$.*
3. *When $\|\mathbf{w}\|_2 \geq c^{\prime}\kappa/(2\epsilon)$, then $\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} \geq c^{\prime}\sqrt{\epsilon}\|\mathbf{w}\|_{\mathbf{w}}$.*

Similarly to the statement of Proposition 14, in our proof we shall consider several cases depending on how large $\|\mathbf{w}\|_2$ is. For the rest of the proof, let $c^{\prime}$ be the absolute constant of Proposition 15 and denote by $\kappa = L^3 R^6 \mu^3 \tau^4/\xi^2$, $\Lambda_1 = 16$, $\Lambda_2 = 1/\kappa \geq \xi^{6/5}/(c^{\prime}\kappa L^2 \mu^6)^{1/5}$ and $K = \Lambda_2^{2.5} U/(\sqrt{c^{\prime}\kappa} LR)$.

First, we show Case 3 of Proposition 14 and bound from below the contribution of the gradient on the direction of $\mathbf{w}$, when the norm of $\mathbf{w}$ is large. In other words, we show that in this case the gradient contribution of the regularizer is large enough so that the gradient field of the regularized L_2^2-objective "pulls" $\mathbf{w}$ towards the origin.

Claim 27 *If $\|\mathbf{w}\|_2 \geq c^{\prime}\kappa/(\epsilon\Lambda_1)$, then $\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} \geq \sqrt{\epsilon}\|\mathbf{w}\|_{\mathbf{w}}$.*

Proof We have that

$$\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} = \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)\sigma^{\prime}(\mathbf{w} \cdot \mathbf{x})\mathbf{w} \cdot \mathbf{x}] + \rho\|\mathbf{w}\|_2^2 .$$

First we bound from below the quantity $\mathbf{E}_{(\mathbf{x},y) \sim D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)\sigma^{\prime}(\mathbf{w} \cdot \mathbf{x})\mathbf{w} \cdot \mathbf{x}]$. We need to bound the maximum value that σ can obtain.

Fact 28 *Let σ be a (τ, μ, ξ)-sigmoidal activation, then $|\sigma(t_1) - \sigma(t_2)| \leq 2\xi/\mu$ for any $t_1, t_2 \in \mathbb{R}$.*

Proof Using the fundamental theorem of calculus, for any $t_1, t_2 \in \mathbb{R}$, it holds

$$|\sigma(t_1) - \sigma(t_2)| = \left|\int_{t_1}^{t_2} \sigma^{\prime}(t)\mathrm{d}t\right| \leq \int_{-1}^{1} \sigma^{\prime}(t)\mathrm{d}t \leq 2\int_0^1 \xi e^{-t\mu}\mathrm{d}t \leq 2\xi/\mu ,$$

where we used that σ is non-decreasing and that $\sigma^{\emptyset}(t) \leq \xi e^{t\mu}$. ■

Using Fact 28, we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\ D}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)\sigma^{\emptyset}(\mathbf{w}\cdot\mathbf{x})\mathbf{w}\cdot\mathbf{x}] &\geq -\mathop{\mathbf{E}}_{(\mathbf{x},y)\ D}[|(\sigma(\mathbf{w}\cdot\mathbf{x})-y)||\sigma^{\emptyset}(\mathbf{w}\cdot\mathbf{x})\mathbf{w}\cdot\mathbf{x}|] \\
&\geq -2\frac{\xi}{\mu}\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[\sigma^{\emptyset}(\mathbf{w}\cdot\mathbf{x})^2|\mathbf{w}\cdot\mathbf{x}|] \\
&\geq -2\frac{\xi^3}{\mu}\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[\exp(-2\mu|\mathbf{w}\cdot\mathbf{x}|\leq 1)|\mathbf{w}\cdot\mathbf{x}|] \\
&\geq -\frac{\xi^3}{L\mu^3\|\mathbf{w}\|_2}\,,
\end{aligned}$$

where in the second inequality we used that the maximal difference of $(\sigma(\mathbf{w}\cdot\mathbf{x})-y)$ is less than $2\xi/\mu$, and in the third the upper bound in the derivative of σ. Therefore, we have that

$$\begin{aligned}
\nabla F_\rho(\mathbf{w})\cdot\mathbf{w} &\geq (1/2)\rho\|\mathbf{w}\|_2^2 + (1/2)\|\mathbf{w}\|_2^2\left(\rho - \frac{2\xi^3}{L\mu^3\|\mathbf{w}\|_2^3}\right) \\
&\geq (1/2)\rho\|\mathbf{w}\|_2^2\,,
\end{aligned}$$

where in the last inequality we used that $\rho \geq \frac{2\epsilon^3\Lambda_1^3\xi^3}{c^{\emptyset 3}\kappa^3\mu^3 L} \geq \frac{2\xi^3}{L\mu^3\boldsymbol{k}\mathbf{w}\boldsymbol{k}_2^3}$.

Therefore, we have that $\nabla F_\rho(\mathbf{w})\cdot\mathbf{w} \geq (1/2)\rho\|\mathbf{w}\|_2^{2.5}\|\mathbf{w}\|_{\mathbf{w}}$. By using that $\|\mathbf{w}\|_2 \geq c^{\emptyset}\kappa/(\epsilon\Lambda_1)$ and $\rho \geq \frac{2\epsilon^3\Lambda_1^3\xi^3}{c^{\emptyset 3}\kappa^3\mu^3 L}$, we get that

$$\nabla F_\rho(\mathbf{w})\cdot\mathbf{w} \geq \sqrt{\epsilon\Lambda_1}\frac{\xi^3}{\mu^3\sqrt{c^{\emptyset}\kappa}L}\|\mathbf{w}\|_{\mathbf{w}} \geq \sqrt{\epsilon}\|\mathbf{w}\|_{\mathbf{w}}\,,$$

where we used that $\xi, \Lambda_1 \geq 1, \mu, \kappa, L \leq 1$ (see Remark 11) and this completes the proof of Claim 27. ■

Next we bound from below the contribution of the gradient in the direction $\mathbf{w} - \mathbf{w}^*$ when $\mathbf{w}$ is close to the origin, i.e., when $\|\mathbf{w}\|_2 \leq 2/R$ (first case of Proposition 14). In this case, we show that our choice of regularization ρ, ensures that when $\mathbf{w}$ is not very large, then the gradient behaves qualitatively similarly to that of the vanilla L_2^2 objective, i.e., as in Proposition 15.

Claim 29 *If $\|\mathbf{w}\|_2 \leq 2/R$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq \sqrt{\epsilon}\xi/(c^{\emptyset}LR^4\tau^2)$, then*

$$\nabla F_\rho(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \geq (c^{\emptyset}/2)\xi\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_2\,.$$

Proof From Proposition 15, we have that if $\|\mathbf{w}\|_2 \leq 2/R$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq \sqrt{\epsilon}\xi/(c^{\emptyset}LR^4\tau^2)$, then $\nabla F(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \geq c^{\emptyset}\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_2$. Therefore, we have that

$$\begin{aligned}
\nabla F_\rho(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) &\geq c^{\emptyset}\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_2 + \rho\|\mathbf{w}\|_2(\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta) \\
&\geq c^{\emptyset}\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_2 - \rho\|\mathbf{w}\|_2\|\mathbf{w} - \mathbf{w}^*\|_2 \\
&\geq (c^{\emptyset}/2)\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_2\,,
\end{aligned}$$

where in the third inequality we used that $\|\mathbf{w}\|_2 \leq 2/R$ and that $\rho = O(F^3(\mathbf{w}^*))$. ■

Next we handle the case where $2/R \leq \|\mathbf{w}\|_2 \leq c'\kappa/(\epsilon\Lambda_2)$ (see Case 2 of Proposition 14). Again, the fact that we choose a small regularization parameter allows us to show that the gradient field in this case behaves similarly to that of the unregularized L_2^2 objective (see Case 2 of Proposition 15).

Claim 30 *If* $2/R \leq \|\mathbf{w}\|_2 \leq c'\kappa/(\epsilon\Lambda_2)$, *then*

$$\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq (\sqrt{c'}/2)\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}\,.$$

Proof We compute the contribution of the regularizer in the direction $\mathbf{w} - \mathbf{w}^*$. This is equal to $\rho\|\mathbf{w}\|_2(\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta)$. Note that this is positive when $\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta \geq 0$ and negative otherwise. Hence, if $\theta \in (\pi/2, \pi)$ the contribution of the regularizer is positive, and therefore it is bounded from below by the contribution of the gradient without the regularizer. We need to choose the value of ρ so that the regularizer cancels out the contribution of the noise when $\|\mathbf{w}\|_2$ is large (i.e., when the regularizer has positive contribution).

From Proposition 15, we have that if $\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}} \geq \sqrt{\epsilon/\kappa}$, then $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c'\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$, hence, it holds

$$\begin{aligned}
\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) &\geq c'\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}} + \rho\|\mathbf{w}\|_2(\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta) \\
&\geq c'\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}} - \rho\|\mathbf{w}\|_2|\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta| \\
&\geq \frac{\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta|}{\|\mathbf{w}\|_2^{3/2}}(c'\sqrt{\epsilon} - \rho\|\mathbf{w}\|_2^{2.5}) + c'\sqrt{\epsilon}\frac{\|\mathbf{w}^*\|_2\sin\theta}{\|\mathbf{w}\|_2} \\
&\geq (c'/2)\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}\,,
\end{aligned}$$

where for the last inequality we used that $\|\mathbf{w}\|_2 \leq c'\kappa/(\epsilon\Lambda_2)$ and that $\rho \leq \frac{\Lambda_2^{2.5}\epsilon^3}{4c'^{1.5}\kappa^{2.5}}$. ■

Finally, we consider the case where $c'\kappa/(\epsilon\Lambda_2) \leq \|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$ and $\|\mathrm{proj}_{\mathbf{w}^*}\mathbf{w}\|_2 \geq 2K/R$. We show that the contribution in the direction $\mathbf{w} - \mathbf{w}^*$ of the unregularized gradient of L_2^2 is greater than the contribution of the gradiend corresponding to the regularizer. The proof of the following claim can be found in Appendix A.

Claim 31 *If* $c'\kappa/(\epsilon\Lambda_2) \leq \|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$ *and* $\|\mathrm{proj}_{\mathbf{w}^*}\mathbf{w}\|_2 \geq 2K/R$ *then*

$$\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq (c'/2)\sqrt{\epsilon}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}\,.$$

To prove the Proposition 14, it remains to show that $\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$ is small when $\|\mathrm{proj}_{\mathbf{w}^*}\mathbf{w}\|_2 \leq 2K/R$ and $c'\kappa/(\epsilon\Lambda_2) \leq \|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$. Similarly to the proof of Claim 31, we need to consider only the case where $\theta \in (0, \pi/2)$ and $\|\mathbf{w}\|_2 \leq 2\|\mathbf{w}^*\|_2$. In fact, we show that in this case the vector $\mathbf{w}$ is close to the target vector $\mathbf{w}^*$. We have:

$$\begin{aligned}
\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}} &\leq \frac{2\|\mathbf{w}\|_2 + \|\mathbf{w}^*\|_2}{\|\mathbf{w}\|_2^{3/2}} + \frac{\|\mathrm{proj}_{\mathbf{w}^*}\mathbf{w}\|_2}{\|\mathbf{w}\|_2^{1/2}} \\
&\leq \frac{K/R + 1}{\|\mathbf{w}\|_2^{1/2}} + \frac{\|\mathbf{w}^*\|_2}{\|\mathbf{w}\|_2^{3/2}}\,,
\end{aligned}$$

where in the first inequality we used the triangle inequality and in the second the fact that $\|\mathrm{proj}_{\mathbf{w}^*}\mathbf{w}\|_2 \leq 2K/R$. Moreover, recall that $c'\kappa/(\epsilon\Lambda_2) \leq \|\mathbf{w}\|_2$, $\|\mathbf{w}^*\|_2 \leq U/\epsilon$ and $K = \Lambda_2^{2.5}U/(\sqrt{c'}\kappa RL)$. Therefore, we have that $\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}} \leq U\sqrt{\epsilon}\Lambda_2^3/(c'\kappa)^{3/2} \leq (cU/\kappa^5)\sqrt{\epsilon}$, for some absolute constant $c > 0$.

A.3. Proof of Theorem 12

We restate and prove the following theorem:

Theorem 32 (Stationary Points of Sigmoidal Activations) *Let D be an ϵ-corrupted, (L,R)-well-behaved distribution on $\mathbb{R}^d \times \mathbb{R}$ and σ be a (τ,μ,ξ)-sigmoidal activation. Set $\kappa = L^6R^6\mu^3\tau^4/\xi^2$ and $\rho = C\epsilon^3/\kappa^5$, where $C>0$ is a universal constant, and define the ℓ_2-regularized objective as $F_\rho(\mathbf{w}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y)\ D}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2] + (\rho/2)\|\mathbf{w}\|_2^2$. For some sufficiently small universal constant $c>0$, we have the following*

- *If $\|\mathbf{w}\|_2 \le 2/R$ and $\|\nabla F_\rho(\mathbf{w})\|_2 \le c\sqrt{\epsilon}$, then $F(\mathbf{w}) = \epsilon\cdot\mathrm{poly}(1/\kappa)$.*
- *If $\|\mathbf{w}\|_2 \ge 2/R$ and $\|\nabla F_\rho(\mathbf{w})\|_{\ ,\mathbf{w}} \le c\sqrt{\epsilon}$, then $F(\mathbf{w}) = \epsilon\cdot\mathrm{poly}(1/\kappa)$.*

Proof We note that from Lemma 13, there exists a $\mathbf{w}^* \in \mathbb{R}^d$, such that $F(\mathbf{w}^*)\|\mathbf{w}^*\|_2 = 1 + O(\xi/(\mu L)) = U$ and $F(\mathbf{w}^*) \le \epsilon$. We can assume that $\epsilon \le \mathrm{poly}(\xi/(LR\mu))$, since otherwise any vector $\mathbf{w}$, gets error $F(\mathbf{w}) \le 2\epsilon$. First we consider the case where $\|\mathbf{w}\|_2 \le 2/R$. From Proposition 14 for $c^\prime > 0$ a sufficiently small absolute constant, we have that if $\|\mathbf{w}-\mathbf{w}^*\|_2 \ge \sqrt{\epsilon}/(c^\prime\kappa^5)$, then

$$\|\nabla F(\mathbf{w})\|_2 \ge \nabla F(\mathbf{w})\cdot\frac{\mathbf{w}-\mathbf{w}^*}{\|\mathbf{w}-\mathbf{w}^*\|_2} \ge c^\prime\sqrt{\epsilon}\,.$$

In order to reach a contradiction, assume that $\|\nabla F(\mathbf{w})\|_2 < c^\prime\sqrt{\epsilon}$ and $\mathbf{E}_{\mathbf{x}\ D_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] > \xi^2\epsilon/(c^{\prime 2}\kappa^{10})$. From Proposition 14, we have that $\|\mathbf{w}-\mathbf{w}^*\|_2 \le \sqrt{\epsilon}/(c^\prime\kappa^5)$. Therefore, from Lemma 43 we have that

$$\xi^2\epsilon/(c^{\prime 2}\kappa^{10}) < \mathop{\mathbf{E}}_{\mathbf{x}\ D_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] \le \|\mathbf{w}-\mathbf{w}^*\|_2^2 \le \xi^2\epsilon/(c^{\prime 2}\kappa^{10})\,,$$

which is a contradiction.

Next, we consider the case where $\|\mathbf{w}\|_2 \ge c^\prime\kappa/(2\epsilon)$. From Proposition 14, we know that there is no approximate stationary point in this region, i.e., there is no point $\mathbf{w}$ with $\|\nabla F(\mathbf{w})\|_{\ ,\mathbf{w}} \le c^\prime\sqrt{\epsilon}$.

For the last case we consider the case where $2/R \le \|\mathbf{w}\|_2 \le c^\prime\kappa/\epsilon$. From Proposition 14, we have that if either $\|\mathbf{w}-\mathbf{w}^*\|_\mathbf{w} \ge U\sqrt{\epsilon}/(c^\prime\kappa^5)$ or $\|\mathbf{w}\|_2 \ge 2\|\mathbf{w}^*\|_2$, then

$$\|\nabla F(\mathbf{w})\|_{\ ,\mathbf{w}} \ge \nabla F(\mathbf{w})\cdot\frac{\mathbf{w}-\mathbf{w}^*}{\|\mathbf{w}-\mathbf{w}^*\|_\mathbf{w}} \ge c^\prime\sqrt{\epsilon}\,,$$

where we used that for any two vectors $\mathbf{z}_1,\mathbf{z}_2\in\mathbb{R}^d$, it holds $\mathbf{z}_1\cdot\mathbf{z}_2 \le \|\mathbf{z}_1\|_{\ ,\mathbf{w}}\|\mathbf{z}_2\|_\mathbf{w}$.

In order to reach a contradiction, assume that $\mathbf{E}_{\mathbf{x}\ D_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] \ge C\frac{\xi^2U^2}{L^4\mu^3\kappa^{10}}\epsilon$ and that $\|\nabla F(\mathbf{w})\|_{\ ,\mathbf{w}} \le c^\prime\sqrt{\epsilon}$, where $C>0$ is a sufficiently large absolute constant. It holds that $\|\mathbf{w}-\mathbf{w}^*\|_\mathbf{w} \le U\sqrt{\epsilon}/(c^\prime\kappa^5)$ and $\|\mathbf{w}\|_2 \le 2\|\mathbf{w}^*\|_2$. Hence, we have that

$$\|\mathbf{w}-\mathbf{w}^*\|_\mathbf{w} = \frac{|\|\mathbf{w}\|_2 - \cos\theta\|\mathbf{w}^*\|_2|}{\|\mathbf{w}\|_2^{3/2}} + \frac{\|\mathbf{w}^*\|_2\sin\theta}{\sqrt{\|\mathbf{w}\|_2}} \le \frac{U\sqrt{\epsilon}}{\kappa^5}\,.$$

Therefore, it holds that

$$\sin\theta \le \frac{\sqrt{\epsilon\|\mathbf{w}\|_2}}{\|\mathbf{w}^*\|_2\kappa^5} \le \frac{\sqrt{\epsilon}}{\sqrt{\|\mathbf{w}^*\|_2}\kappa^5}\,,$$

where we used that $\|\mathbf{w}\|_2 \leq 2\|\mathbf{w}\ \|_2$, and using that $\sqrt{\epsilon/\|\mathbf{w}\ \|_2}$ is smaller than a sufficiently small absolute constant, we get that $\sin\theta \leq 1/2$, and hence, $\theta \in (0, \pi/4)$. Using again Lemma 43, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}\ \cdot\mathbf{x}))^2]\ .\ \frac{\xi^2}{L^4\mu^3}\|\mathbf{w}-\mathbf{w}\ \|_{\mathbf{w}}^2\ .\ \frac{\xi^2U^2}{L^4\mu^3\kappa^{10}}\epsilon\,.$$

Therefore, we get again a contradiction. The proof then follows by noting that $F(\mathbf{w}) \leq 2\epsilon + 2\,\mathbf{E}_{\mathbf{x}\ D_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}\ \cdot\mathbf{x}))^2]$.

■

A.4. Proof of Proposition 15

We restate and prove the following proposition.

Proposition 33 (Gradient of the "Vanilla" L_2^2 Loss (Inside a Ball)) *Let D be an (L,R)-well-behaved distribution and let σ be a (τ,μ,ξ)-sigmoidal activation. Let $\epsilon \in (0,1)$ and let $\mathbf{w} \in \mathbb{R}^d$ with $F(\mathbf{w}\ ) \leq \epsilon$ and $F(\mathbf{w}\ )$ is less than a sufficiently small multiple of $L^2R^6\tau^4/\xi^2$. Denote $\kappa = L^6R^6\mu^3\tau^4/\xi^2$. There exists a universal constant $c^{\emptyset} > 0$, such that for any $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq c^{\emptyset}\kappa/\epsilon$, it holds that*

1. *When $\|\mathbf{w}\|_2 \leq 2/R$ and $\|\mathbf{w}\ -\mathbf{w}\|_2 \geq \sqrt{\epsilon}\xi/(c^{\emptyset}LR^4\tau^2)$, then*
$$\nabla F(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}\ ) \geq c^{\emptyset}\sqrt{\epsilon}\|\mathbf{w}\ -\mathbf{w}\|_2\,.$$
2. *When $\|\mathbf{w}\|_2 \geq 2/R$ and either $\|\mathbf{w}\ -\mathbf{w}\|_{\mathbf{w}} \geq \sqrt{\epsilon/\kappa}$ or $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}\ \|_2$, then*
$$\nabla F(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}\ ) \geq c^{\emptyset}\sqrt{\epsilon}\|\mathbf{w}\ -\mathbf{w}\|_{\mathbf{w}}\,.$$

Proof We decompose the gradient into a part that corresponds to the contribution of the "true" labels, i.e., $\sigma(\mathbf{w}\ \cdot\mathbf{x})$ (see I_2 below), and a part corresponding to the noise (see I_1 below). We have that

$$\begin{aligned}\nabla F(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}\ ) = &\underbrace{\mathbf{E}[(\sigma(\mathbf{w}\ \cdot\mathbf{x}) - y)\sigma^{\emptyset}(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}\ \cdot\mathbf{x})]}_{I_1}\\ &+\underbrace{\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}\ \cdot\mathbf{x}))\sigma^{\emptyset}(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}\ \cdot\mathbf{x})]}_{I_2}\,. \qquad (4)\end{aligned}$$

In the following subsections, we bound I_1 and I_2 from below. We start by bounding from below the effect of the noise, i.e., the contribution of I_1.

Estimating the Effect of the Noise We start by showing that the noise cannot affect the gradient by a lot, i.e., we bound the contribution of I_1. We prove the following lemma:

Lemma 34 *Let D be an (L,R)-well-behaved distribution and σ be a (τ,μ,ξ)-sigmoidal activation. For any vector $\mathbf{w} \in \mathbb{R}^d$, it holds that*

$$I_1 \geq -\sqrt{8\xi^2\epsilon}\min(\|\mathbf{w}\ -\mathbf{w}\|_{\mathbf{w}}/(\mu^{3/2}L^2), \|\mathbf{w}\ -\mathbf{w}\|_2)\,.$$

Proof Using the Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned} I_1 &\geq -\mathbf{E}[|(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})|] \\ &\geq -(\mathbf{E}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2])^{1/2}(\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2\sigma'(\mathbf{w} \cdot \mathbf{x})^2])^{1/2} \\ &= -\sqrt{F(\mathbf{w}^*)}\,(\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2\sigma'(\mathbf{w} \cdot \mathbf{x})^2])^{1/2} \\ &\geq -\sqrt{\epsilon}\,(\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2\sigma'(\mathbf{w} \cdot \mathbf{x})^2])^{1/2}\,, \end{aligned} \tag{5}$$

where we used that $F(\mathbf{w}^*) \leq \epsilon$, from the assumptions of Proposition 15. We proceed to bound the term $\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2\sigma'(\mathbf{w} \cdot \mathbf{x})^2]$. Note that we can use the upper bound on the derivative of the activation function, i.e., $\sigma'(t) \leq \xi$ for all $t \in \mathbb{R}$. However, this would result in $\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2\sigma'(\mathbf{w} \cdot \mathbf{x})^2] \leq O(\xi^2\|\mathbf{w} - \mathbf{w}^*\|_2^2)$. While, this was sufficient for the case of unbounded activations of the Definition 6, for bounded activation functions, we need a tighter estimate that takes into account the fact that the functions have exponential tails outside the interval $[-1/\mu, +1/\mu]$. Recall that we denote by $\mathbf{q}^{\parallel \mathbf{w}^*}$ the component of $\mathbf{q} \in \mathbb{R}^d$ parallel to $\mathbf{w}^* \in \mathbb{R}^d$, i.e., $\mathbf{q}^{\parallel \mathbf{w}^*} = \mathrm{proj}_{\mathbf{w}^*}\,\mathbf{q}$. Similarly, we denote $\mathbf{q}^{\perp \mathbf{w}^*} = \mathrm{proj}_{\mathbf{w}^{*\perp}}\mathbf{q}$. We prove the following.

Lemma 35 *Let D be an (L, R)-well-behaved distribution and σ be a (τ, μ, ξ)-sigmoidal activation. For any vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, it holds that*

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{v} \cdot \mathbf{x})^2(\sigma'(\mathbf{w} \cdot \mathbf{x}))^2] \leq 8\xi^2 \min\left(\frac{1}{L^4\mu^3}\|\mathbf{v} - \mathbf{w}\|_{\mathbf{w}}^2, \|\mathbf{v} - \mathbf{w}\|_2^2\right)\,.$$

Proof First note that $\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x}-\mathbf{v}\cdot\mathbf{x})^2\sigma'(\mathbf{w}\cdot\mathbf{x})^2] \leq \xi^2\,\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x}-\mathbf{v}\cdot\mathbf{x})^2\exp(-2|\mathbf{w}\cdot\mathbf{x}|\mu)]$, from the assumption that $|\sigma'(t)| \leq \xi\exp(-\mu|t|)$. It holds that

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{v} \cdot \mathbf{x})^2\exp(-2|\mathbf{w} \cdot \mathbf{x}|\mu)] \leq \mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{v} \cdot \mathbf{x})^2] = \|\mathbf{w} - \mathbf{v}\|_2^2\,,$$

where we used the fact that the distribution $D_{\mathbf{x}}$ is isotropic, i.e., for any vector $\mathbf{u} \in \mathbb{R}^d$, it holds that $\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{u} \cdot \mathbf{x})^2] = \|\mathbf{u}\|_2^2$.

Next show that by using the tails of the distribution $D_{\mathbf{x}}$, we can prove a tighter upper bound for some cases. We use that the distribution is (L, R)-well-behaved, and we prove the following claim that bounds from above the expectation.

Claim 36 *Let $D_{\mathbf{x}}$ be an (L, R)-well-behaved distribution. Let $b > 0$ and $\mathbf{u}, \mathbf{v}$ be unit norm orthogonal vectors. It holds that*

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{u} \cdot \mathbf{x})^2\exp(-b|\mathbf{v} \cdot \mathbf{x}|)] \leq \frac{8}{L^4 b} \quad \mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{v} \cdot \mathbf{x})^2\exp(-b|\mathbf{v} \cdot \mathbf{x}|)] \leq \frac{8}{L^2 b^3}\,.$$

Proof Without loss of generality take $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{v} = \mathbf{e}_2$. Using the fact that the distribution is (L, R)-well-behaved, we have that the 2-dimensional projection on the subspace V spanned by $\mathbf{v}, \mathbf{u}$ is bounded from above by $(1/L)\exp(-L\|\mathbf{x}_V\|_2)$ everywhere. We have that

$$\begin{aligned} \mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[\mathbf{x}_1^2\exp(-b\mathbf{v} \cdot \mathbf{x})] &\leq (1/L)\int_{\mathbf{x}_2 \in \mathbb{R}}\int_{\mathbf{x}_1 \in \mathbb{R}} \mathbf{x}_1^2\exp(-b\mathbf{v} \cdot \mathbf{x})\exp(-L\|\mathbf{x}_V\|_2)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2 \\ &\leq (4/L)\int_0^1 \mathbf{x}_1^2\exp(-L|\mathbf{x}_1|)\int_0^1 \exp(-b|\mathbf{x}_2|)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2 \\ &= 8/(L^4 b)\,. \end{aligned}$$

Putting the above estimates together, we proved the first part of Claim 36. For the other part, it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{v}\cdot\mathbf{x})^2\exp(-b\mathbf{v}\cdot\mathbf{x})] \le (4/L)\int_0^1\int_0^1 \mathbf{x}_2^2\exp(-L|\mathbf{x}_1|)\exp(-b|\mathbf{x}_2|)\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{x}_1 = 8/(L^2b^3)\ .$$

This completes the proof of Claim 36. ■

Let $\mathbf{q}=\mathbf{v}-\mathbf{w}$ and note that $(\mathbf{q}\cdot\mathbf{x})^2=(\mathbf{q}^{k_\mathbf{w}}\cdot\mathbf{x})^2+(\mathbf{q}^{?_\mathbf{w}}\cdot\mathbf{x})^2$. Using Claim 36, it holds that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{q}\cdot\mathbf{x})^2\exp(-2\mathbf{w}\cdot\mathbf{x}\mu)] &= \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{q}^{k_\mathbf{w}}\cdot\mathbf{x})^2+(\mathbf{q}^{?_\mathbf{w}}\cdot\mathbf{x})^2)\exp(-2\mathbf{w}\cdot\mathbf{x}\mu)]\\ &\le \frac{8}{L^4}(\|\mathbf{q}^{k_\mathbf{w}}\|_2^2\frac{1}{\mu^3\|\mathbf{w}\|_2^3}+\|\mathbf{q}^{?_\mathbf{w}}\|_2^2\frac{1}{\mu\|\mathbf{w}\|_2}) \le \frac{8\|\mathbf{q}\|_\mathbf{w}^2}{L^4\mu^3}\ ,\end{aligned}$$

where we used that $a^2+b^2\le(a+b)^2$ for $a,b\ge 0$. This completes the proof of Lemma 35. ■

Lemma 34 follows from combining Equation (5) along with Lemma 35. ■

Estimating the Contribution of the "Noiseless" Gradient In order to compute the contribution of the gradient when there is no noise to the instance, we show that in fact the contribution of I_2 is bounded from below by the contribution of a ramp function instead of σ. To show this, we use the property that $\sigma^{\emptyset}(t)\ge\tau$, for all $t\in[-1,1]$, see Definition 3.

Claim 37 *It holds that* $I_2\ge\tau^2\,\mathbf{E}_{\mathbf{x}\ D_{\mathbf{x}}}[(r(\mathbf{w}\cdot\mathbf{x})-r(\mathbf{v}\cdot\mathbf{x}))r^{\emptyset}(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}]$.

Proof We have that

$$\begin{aligned}I_2 &= \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))\sigma^{\emptyset}(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}]\\ &= \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[|(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))\sigma^{\emptyset}(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}|]\\ &\ge \tau E_{\mathbf{x}\ D_{\mathbf{x}}}[|(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}|\mathtt{1}\{|\mathbf{w}\cdot\mathbf{x}|\le 1\}]\ ,\end{aligned}$$

where we used that σ is non-decreasing and that $\sigma(t)\ge\tau$ for $t\in[-1,1]$ from the assumptions of (τ,μ,ξ)-sigmoidal activations. Moreover, note that from the fundamental theorem of calculus, we have that

$$\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}) = \int_{\mathbf{v}\cdot\mathbf{x}}^{\mathbf{w}\cdot\mathbf{x}}\sigma^{\emptyset}(t)\mathrm{d}t \ge \tau\int_{\mathbf{v}\cdot\mathbf{x}}^{\mathbf{w}\cdot\mathbf{x}}\mathtt{1}\{|t|\le 1\}\mathrm{d}t = \tau(r(\mathbf{w}\cdot\mathbf{x})-r(\mathbf{v}\cdot\mathbf{x}))\ ,$$

where we used again that $\sigma(t)\ge\tau$ for $t\in[-1,1]$. Therefore, we have that

$$I_2\ge\tau^2\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(r(\mathbf{w}\cdot\mathbf{x})-r(\mathbf{v}\cdot\mathbf{x}))r^{\emptyset}(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}]\ .$$

This completes the proof of Claim 37. ■

To bound I_2 from below, we consider three cases depending on how far the vector $\mathbf{w}$ is from the target vector $\mathbf{w}^*$. The first two cases correspond to $\theta(\mathbf{w},\mathbf{w}^*)\in(0,\pi/2)$. In the first case, we have that either $\|\mathrm{proj}_{\mathbf{w}^{?_\mathbf{w}}}\mathbf{w}^*\|_2\ge 2K/R$, for any $K\ge 1$, or $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2\ge 2\|\mathbf{w}\|_2$ and in

the second case when both of the aforementioned conditions are false. The last case corresponds to $\theta(\mathbf{w}, \mathbf{w}^*) \in (\pi/2, \pi)$. Notice that when $\mathbf{w}$ is close to the target $\mathbf{w}^*$ then its projection onto the orthogonal complement of $\mathbf{w}$, i.e., $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2$, will be small and also its projection on the direction of $\mathbf{w}$, i.e., $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2$, will be close to $\|\mathbf{w}\|_2$.

If either of the conditions of the first case are satisfied or when $\mathbf{w} \cdot \mathbf{w}^* \leq 0$, then there is a large enough region that we can substitute the $r(\mathbf{w}^* \cdot \mathbf{x})$ by the constant function $r(\mathbf{w}^* \cdot \mathbf{x}) = 1$ (left plot in Figure 3), these conditions corresponds to the case that the vectors $\mathbf{w}$ and $\mathbf{w}^*$ are "far" apart from each other. Whereas if both of these conditions are not satisfied then there exists a large enough region that $r(\mathbf{w}^* \cdot \mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x}$ (right plot in Figure 3), in this case the vectors $\mathbf{w}$ and $\mathbf{w}^*$ are "close" to each other. Without loss of generality, we may assume that $\mathbf{w}/\|\mathbf{w}\|_2 = \mathbf{e}_2$ and $\mathbf{w}^* = \|\mathbf{w}^*\|_2(\cos\theta\mathbf{e}_2 - \sin\theta\mathbf{e}_1)$. For simplicity, we abuse notation and denote by $D_{\mathbf{x}}$ the marginal distribution on the subspace spanned by the vectors $\mathbf{w}, \mathbf{w}^*$.

Case 1: $\mathbf{w}$ and $\mathbf{w}^*$ are "close" We now handle the case where $\theta(\mathbf{w}, \mathbf{w}^*) \in (0, \pi/2)$, $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \leq 2K/R$ and $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2 \leq 2\|\mathbf{w}\|_2$, i.e., the case where $\mathbf{w}$ and $\mathbf{w}^*$ are close to each other but still not close enough to guarantee small L^2 error, i.e., the L_2^2 error of $\mathbf{w}$, $F(\mathbf{w})$ is much larger than ϵ.

Lemma 38 *Let D be an (L, R)-well-behaved distribution. For any vector $\mathbf{w} \in \mathbb{R}^d$, let $\theta = \theta(\mathbf{w}^*, \mathbf{w}) \in (0, \pi/2)$. If $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \leq 2K/R$, for $K \geq 1$, and $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2 \leq 2\|\mathbf{w}\|_2$, it holds:*

- *If $\|\mathbf{w}\|_2 \geq 2/R$, then $I_2 \geq \frac{\tau^2 LR^3}{4096K^3}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}^2$.*
- *Otherwise, $I_2 \geq \frac{\tau^2 LR^4}{2048K^3}\|\mathbf{w} - \mathbf{w}^*\|_2^2$.*

Proof Using Claim 37, we have that

$$\begin{aligned} I_2 &\geq \tau^2 \mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(r(\mathbf{w} \cdot \mathbf{x}) - r(\mathbf{w}^* \cdot \mathbf{x}))r'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} - \mathbf{w}^*) \cdot \mathbf{x}] \\ &\geq \tau^2 \mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}[(\mathbf{w} \cdot \mathbf{x} - r(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{w} - \mathbf{w}^*) \cdot \mathbf{x}\mathbf{1}\{0 \leq \mathbf{w} \cdot \mathbf{x} \leq 1\}] \,. \end{aligned}$$

Note that in the last inequality we used that because $r(\cdot)$ is a non-decreasing function, it holds that $(r(\mathbf{w} \cdot \mathbf{x}) - r(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{w} - \mathbf{w}^*) \cdot \mathbf{x} \geq 0$ for any value of $\mathbf{x} \in \mathbb{R}^d$. Moreover, we assume without loss of generality that $\mathbf{w}/\|\mathbf{w}\|_2 = \mathbf{e}_2$, and therefore $\mathbf{w}^* = \|\mathbf{w}^*\|_2(\cos\theta\mathbf{e}_2 - \sin\theta\mathbf{e}_1)$. Note that in this case the condition $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2 \leq 2\|\mathbf{w}\|_2$ is equivalent to $2\|\mathbf{w}\|_2 \geq |\cos\theta|\|\mathbf{w}^*\|_2$.

Consider the region $-R/(4K) \leq \mathbf{x}_1 \leq -R/(8K)$ and $\mathbf{x}_2 \leq 1/(4\|\mathbf{w}\|_2)$ which is chosen to guarantee that $0 \leq \mathbf{w}^* \cdot \mathbf{x} \leq 1$, which holds because

$$0 \leq \mathbf{w}^* \cdot \mathbf{x} \leq \|\mathbf{w}^*\|_2 \cos\theta\mathbf{x}_2 + R\|\mathbf{w}^*\|_2 \sin\theta/(4K) \leq 1/2 + \cos\theta\|\mathbf{w}^*\|_2/(4\|\mathbf{w}\|_2) \leq 1 \,.$$

We show the following claim which will be used to bound from below I_2 for this case.

Claim 39 *Let $D_{\mathbf{x}}$ be an (L, R)-well-behaved distribution. For any vectors $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$ and $a, b \in [0, R]$, let $\mathbf{q} = \mathbf{w} - \mathbf{v}$, it holds*

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}\left[(\mathbf{q} \cdot \mathbf{x})^2 \mathbf{1}\left\{\frac{\mathbf{w}}{\|\mathbf{w}\|_2} \cdot \mathbf{x} \in (0, a), \frac{\mathbf{v}^{\mathrm{proj}_{\mathbf{w}}}}{\|\mathbf{v}^{\mathrm{proj}_{\mathbf{w}}}\|_2} \cdot \mathbf{x} \in (-b, 0)\right\}\right] \geq \frac{Lab}{8}(\|\mathbf{q}^{k_{\mathbf{w}}}\|_2^2 a^2 + \|\mathbf{q}^{\perp_{\mathbf{w}}}\|_2^2 b^2) \,.$$

Proof We have that $\mathbf{q} = \mathbf{q}^{k_\mathbf{w}} + \mathbf{q}^{?_\mathbf{w}}$ and using the Pythagorean theorem we have that $((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2 = \|\mathbf{q}^{k_\mathbf{w}}\|_2^2(\mathbf{e}_1\cdot\mathbf{x})^2 + \|\mathbf{q}^{?_\mathbf{w}}\|_2^2(\mathbf{e}_2\cdot\mathbf{x})^2$. Therefore, we have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\ D_\mathbf{x}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2 \mathtt{1}\{\mathbf{e}_1\cdot\mathbf{x}\in(0,a), \mathbf{e}_2\cdot\mathbf{x}\in(-b,0)\}] \\
&\quad\geq \mathop{\mathbf{E}}_{\mathbf{x}\ D_\mathbf{x}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2 \mathtt{1}\{\mathbf{e}_1\cdot\mathbf{x}\in(\frac{a}{2},a), \mathbf{e}_2\cdot\mathbf{x}\in(-b,-\frac{b}{2})\}] \\
&\quad\geq \mathop{\mathbf{E}}_{\mathbf{x}\ D_\mathbf{x}}[(\|\mathbf{q}^{k_\mathbf{w}}\|_2^2(\mathbf{e}_1\cdot\mathbf{x})^2 + \|\mathbf{q}^{?_\mathbf{w}}\|_2^2(\mathbf{e}_2\cdot\mathbf{x})^2) \mathtt{1}\{\mathbf{e}_1\cdot\mathbf{x}\in(\frac{a}{2},a), \mathbf{e}_2\cdot\mathbf{x}\in(-b,-\frac{b}{2})\}] \\
&\quad\geq \frac{1}{4}(\|\mathbf{q}^{k_\mathbf{w}}\|_2^2 a^2 + \|\mathbf{q}^{?_\mathbf{w}}\|_2^2 b^2 \mathop{\mathbf{E}}_{\mathbf{x}\ D_\mathbf{x}}[\mathtt{1}\{\mathbf{e}_1\cdot\mathbf{x}\in(\frac{a}{2},a), \mathbf{e}_2\cdot\mathbf{x}\in(-b,-\frac{b}{2}\}] \\
&\quad\geq \frac{Lab}{8}(\|\mathbf{q}^{k_\mathbf{w}}\|_2^2 a^2 + \|\mathbf{q}^{?_\mathbf{w}}\|_2^2 b^2)\ ,
\end{aligned}$$

where in the last inequality we used that the distribution $D_\mathbf{x}$ is (L,R)-well-behaved. ■

For the case where $\|\mathbf{w}\|_2 \geq 2/R$. Using the Claim 39, we have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\ D_\mathbf{x}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2 \mathtt{1}\{0\leq\mathbf{x}_2\leq 1/(4\|\mathbf{w}\|_2), -R/(4K)\leq\mathbf{x}_1\leq -R/(8K)\}] \\
&\qquad\geq \frac{LR}{128K\|\mathbf{w}\|_2}\left(\|\mathbf{q}^{k_\mathbf{w}}\|_2^2\frac{1}{16\|\mathbf{w}\|_2^2} + \|\mathbf{q}^{?_\mathbf{w}}\|_2^2\frac{R^2}{K^2 16}\right) \\
&\qquad\geq \frac{LR^3}{K^3 2048}\left(\frac{\|\mathbf{q}^{k_\mathbf{w}}\|_2^2}{\|\mathbf{w}\|_2^3} + \frac{\|\mathbf{q}^{?_\mathbf{w}}\|_2^2}{\|\mathbf{w}\|_2}\right) \geq \frac{LR^3}{4096K^3}\|\mathbf{q}\|_\mathbf{w}^2\ .
\end{aligned}$$

For the case where $\|\mathbf{w}\|_2 \leq 2/R$, using Claim 39, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\ D_\mathbf{x}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2 \mathtt{1}\{0\leq\mathbf{x}_2\leq 1/(4\|\mathbf{w}\|_2), -R/(4K)\leq\mathbf{x}_1\leq -R/(8K)\}] \\
&\qquad\geq \mathop{\mathbf{E}}_{\mathbf{x}\ D_\mathbf{x}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2 \mathtt{1}\{R/8\leq\mathbf{x}_2\leq R/4, -R/(4K)\leq\mathbf{x}_1\leq -R/(8K)\}] \\
&\qquad\geq \frac{LR^2}{K128}\left(\|\mathbf{q}^{k_\mathbf{w}}\|_2^2\frac{R^2}{16} + \|\mathbf{q}^{?_\mathbf{w}}\|_2^2\frac{R^2}{K^2 16}\right) \geq \frac{LR^4}{K^3 2048}\|\mathbf{q}\|_2^2\ .
\end{aligned}$$

This completes the proof of Lemma 38. ■

From Lemma 34, we have that $I_1 \geq -\sqrt{8\epsilon/\mu^3}(\xi/L^2)\|\mathbf{w}\ -\mathbf{w}\|_\mathbf{w}$. Using Lemma 38, for $K=1$, we have that

$$\begin{aligned}
I_1 + I_2 &\geq \frac{\tau^2 LR^3}{4096}\|\mathbf{w}\ -\mathbf{w}\|_\mathbf{w}\left(\|\mathbf{w}\ -\mathbf{w}\|_\mathbf{w} - \sqrt{\epsilon}\frac{4096\sqrt{8}\xi}{L^3\mu^{3/2}R^3\tau^2}\right) \\
&\geq 2(\xi/\mu^{3/2})\sqrt{8\epsilon}\|\mathbf{w}\ -\mathbf{w}\|_\mathbf{w} \geq \sqrt{\epsilon}\|\mathbf{w}\ -\mathbf{w}\|_\mathbf{w}\ ,
\end{aligned}$$

where in the second inequality we used that $\|\mathbf{w}\ -\mathbf{w}\|_\mathbf{w} \geq \sqrt{\epsilon}\frac{8192^{\bar{p}}\xi}{L^3\mu^{3/2}R^3\tau^2}$ and that $\xi\geq 1, \mu\leq 1$. Moreover, note that in the special case that $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}\ \|_2$, we have that $\|\mathbf{w}\ -\mathbf{w}\|_\mathbf{w} \geq$

$(1/2)\|\mathbf{w}\|_2{}^{1/2}$. Therefore, it holds

$$\begin{aligned} I_1 + I_2 &\geq \frac{\tau^2 LR^3}{4096}\|\mathbf{w} \ - \mathbf{w}\|_{\mathbf{w}} \left(\|\mathbf{w} \ - \mathbf{w}\|_{\mathbf{w}} - \sqrt{\epsilon}\frac{4096\sqrt{8}\xi}{L^3\mu^{3/2}R^3\tau^2} \right) \\ &\geq \frac{\tau^2 LR^3}{4096}\|\mathbf{w} \ - \mathbf{w}\|_{\mathbf{w}} \left((1/2)\|\mathbf{w}\|_2{}^{1/2} - \sqrt{\epsilon}\frac{4096\sqrt{8}\xi}{L^3\mu^{3/2}R^3\tau^2} \right) \\ &\geq \frac{\tau^2 LR^3}{16384}\frac{\|\mathbf{w} \ - \mathbf{w}\|_{\mathbf{w}}}{\|\mathbf{w}\|_2^{1/2}} \ , \end{aligned}$$

where in the third inequality we used that $\epsilon\|\mathbf{w}\|_2\xi^2/(L^6R^6\mu^3\tau^2)$ is less than a sufficient small constant. Using that $\|\mathbf{w}\|_2 \leq CL^6R^6\mu^3\tau^4/(\epsilon\xi^2)$ for a sufficiently small constant $C > 0$, from the assumptions of Proposition 15, we get that $I_1 + I_2 \geq c^{\prime\prime}\frac{\xi}{\mu^{3/2}}\sqrt{\epsilon}\|\mathbf{w} \ - \mathbf{w}\|_{\mathbf{w}} \geq c^{\prime\prime}\sqrt{\epsilon}\|\mathbf{w} \ - \mathbf{w}\|_{\mathbf{w}}$, where we used that $\xi \geq 1, 1 \geq \mu$ by assumption.

Moreover, for the case that $\|\mathbf{w}\|_2 \leq 2/R$, we have $I_1 \geq -\sqrt{\epsilon}\xi\|\mathbf{w} \ - \mathbf{w}\|_2$. Using Lemma 38, we have that

$$\begin{aligned} I_1 + I_2 &\geq \frac{\tau^2 LR^4}{2048}\|\mathbf{w} \ - \mathbf{w}\|_2 \left(\|\mathbf{w} \ - \mathbf{w}\|_2 - \sqrt{\epsilon}\frac{2048\xi}{LR^4\tau^2} \right) \\ &\geq \sqrt{\epsilon}\|\mathbf{w} \ - \mathbf{w}\|_2 \ , \end{aligned}$$

where in the last inequality we used that $\|\mathbf{w} \ - \mathbf{w}\|_2 \geq \sqrt{\epsilon}\frac{4096\xi}{LR^4\tau^2}$ and that $\xi \geq 1$.

Case 2: $\mathbf{w}$ is far from the target $\mathbf{w}$ We now handle the case where our current guess $\mathbf{w}$ is far from the target weight vector $\mathbf{v}$. In particular, we assume that either $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w} \ \|_2 \geq 2K/R$ for some $K \geq 1$, or $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w} \ \|_2 \geq 2\|\mathbf{w}\|_2$. In this case there is a large enough region for which we can substitute the $r(\mathbf{w} \ \cdot \mathbf{x})$ by the constant function $r(\mathbf{w} \ \cdot \mathbf{x}) = 1$ (left plot in Figure 3). First, we handle the case where $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w} \ \|_2 \geq 2K/R$ or $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w} \ \|_2 \geq 2\|\mathbf{w}\|_2$. We prove the following:

Lemma 40 *Let D be an (L, R)-well-behaved distribution. Let $\theta = \theta(\mathbf{w} \ , \mathbf{w}) \in (0, \pi/2)$. There exists a sufficiently small universal constant $c^{\prime\prime} > 0$ such that for any $\mathbf{w} \in \mathbb{R}^d$, if $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w} \ \|_2 \geq 2K/R$, for $K \geq 1$, or $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w} \ \|_2 \geq 2\|\mathbf{w}\|_2$, it holds:*

- *If $\|\mathbf{w}\|_2 \geq 2/R$, then $I_2 \geq \frac{\tau^2 LR^2}{72k\|\mathbf{w}\|_2^{1/2}}\|\mathbf{w} - \mathbf{w} \ \|_{\mathbf{w}}$.*
- *Otherwise, $I_2 \geq \frac{\tau^2 LR^3}{144}\|\mathbf{w} - \mathbf{w} \ \|_2$.*

Proof From Claim 37, we have that

$$\begin{aligned} I_2 \geq \tau^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D_{\mathbf{x}}}[(r(\mathbf{w}\cdot\mathbf{x}) - r(\mathbf{w} \ \cdot\mathbf{x}))r^{\prime\prime}(\mathbf{w}\cdot\mathbf{x})(\mathbf{w} - \mathbf{w} \ )\cdot\mathbf{x}] \\ = \tau^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w} \ \cdot\mathbf{x}))(\mathbf{w} - \mathbf{w} \ )\cdot\mathbf{x}\,\mathbf{1}\{|\mathbf{w}\cdot\mathbf{x}| \leq 1\}] \\ \geq \tau^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w} \ \cdot\mathbf{x}))(\mathbf{w} - \mathbf{w} \ )\cdot\mathbf{x}\,\mathbf{1}\{0 \leq \mathbf{w}\cdot\mathbf{x} \leq 1\}] \ . \end{aligned}$$

Note that the last inequality holds because r is a non-decreasing function, and thus we have that $(r(\mathbf{w}\cdot\mathbf{x}) - r(\mathbf{w} \ \cdot\mathbf{x}))(\mathbf{w}\cdot\mathbf{x} - \mathbf{w} \ \cdot\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$. Let $c = \min(1/\|\mathbf{w}\|_2, R)$, and note that the

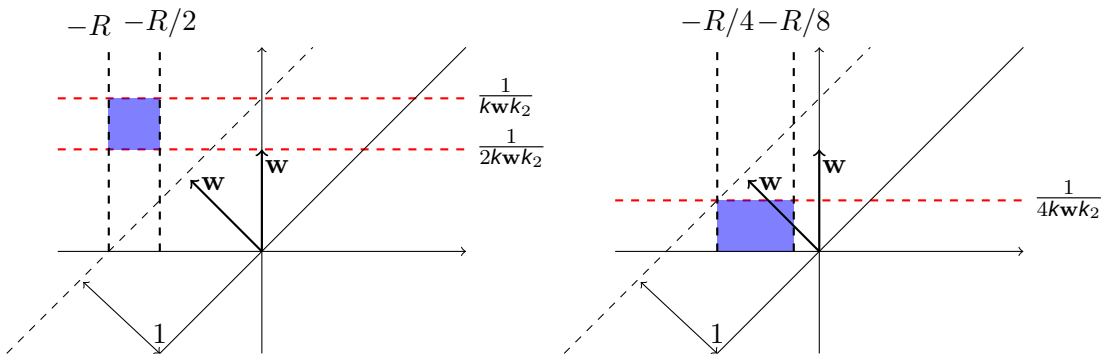


Figure 3: Using our distributional assumptions, there exists a region ("blue") that provides enough contribution to the gradient, the left plot corresponds to Lemma 40 where the angle between the current vector and the target one is large, in this case we take a region where $r(\mathbf{w} \cdot \mathbf{x}) = 1$, the right plot corresponds to Lemma 38, in this case the angle is small, and we take a region where $r(\mathbf{w} \cdot \mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$.

condition $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}\|_2 \geq 2\|\mathbf{w}\|_2$ is equivalent to $2\|\mathbf{w}\| \leq \cos\theta\|\mathbf{w}\|_2$. We consider the following subset $-R \leq \mathbf{x}_1 \leq -R/2$ and $c/2 \leq \mathbf{x}_2 \leq c$ which is chosen such $\mathbf{w} \cdot \mathbf{x} \geq 1$; which holds because $\mathbf{w} \cdot \mathbf{x} \geq \|\mathbf{w}\|_2(\cos\theta\mathbf{x}_2 + R\sin\theta/2) \geq \|\mathbf{w}\|_2(\cos\theta c/2 + R\sin\theta/2)$ and the last part is always greater than 1 if at least one of the following hold: $\|\mathbf{w}\|_2 \sin\theta \geq K/R$ or $2\|\mathbf{w}\| \leq \cos\theta\|\mathbf{w}\|_2$. Therefore, it holds that $r(\mathbf{w} \cdot \mathbf{x}) = 1$. Hence, we can write

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x} D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w}\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w})\cdot\mathbf{x}\,\mathbf{1}\{0 \leq \mathbf{w}\cdot\mathbf{x} \leq 1\}] \\
&\quad\geq \mathop{\mathbf{E}}_{\mathbf{x} D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w}\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w})\cdot\mathbf{x}\,\mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq c, -R \leq \mathbf{x}_1 \leq -R/2\}] \\
&\quad\geq \mathop{\mathbf{E}}_{\mathbf{x} D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - 1)(\mathbf{w}-\mathbf{w})\cdot\mathbf{x}\,\mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq 2c/3, -R \leq \mathbf{x}_1 \leq -R/2\}]\ .
\end{aligned}$$

Notice that in this case $\mathbf{w}\cdot\mathbf{x} \leq \mathbf{w}\cdot\mathbf{x}$ and that $(1 - \mathbf{w}\cdot\mathbf{x}) \geq 1/3$. Therefore,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x} D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w}\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w})\cdot\mathbf{x}\,\mathbf{1}\{0 \leq \mathbf{w}\cdot\mathbf{x} \leq 1\}] \\
&\quad\geq (1/3)\mathop{\mathbf{E}}_{\mathbf{x} D_{\mathbf{x}}}[|(\mathbf{w}-\mathbf{w})\cdot\mathbf{x}|\mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq 2c/3, -R \leq \mathbf{x}_1 \leq -R/2\}]\ .
\end{aligned}$$

Let $\mathbf{q} = \mathbf{w} - \mathbf{w}$. By using the inequality $(a^2+b^2)^{1/2} \geq (1/2)(|a|+|b|)$ for any $a, b \in \mathbb{R}$ and that $(\mathbf{q}\cdot\mathbf{x})^2 = (\mathbf{q}^{k\mathbf{w}}\cdot\mathbf{x})^2 + (\mathbf{q}^{?\mathbf{w}}\cdot\mathbf{x})^2$, we have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x} D_{\mathbf{x}}}[|(\mathbf{w}-\mathbf{w})\cdot\mathbf{x}|\mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq 2c/3, -R \leq \mathbf{x}_1 \leq -R/2\}] \\
&\quad= \mathop{\mathbf{E}}_{\mathbf{x} D_{\mathbf{x}}}[\sqrt{(\mathbf{q}\cdot\mathbf{x})^2}\mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq 2c/3, -R \leq \mathbf{x}_1 \leq -R/2\}] \\
&\quad\geq \frac{1}{2}\mathop{\mathbf{E}}_{\mathbf{x} D_{\mathbf{x}}}[(|\mathbf{q}^{k\mathbf{w}}\cdot\mathbf{x}| + |\mathbf{q}^{?\mathbf{w}}\cdot\mathbf{x}|)\mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq 2c/3, -R \leq \mathbf{x}_1 \leq -R/2\}] \\
&\quad\geq \frac{1}{2}\left(\|\mathbf{q}^{k\mathbf{w}}\|_2\frac{c}{2} + \|\mathbf{q}^{?\mathbf{w}}\|_2\frac{R}{2}\right)\mathop{\mathbf{E}}_{\mathbf{x} D_{\mathbf{x}}}[\mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq 2c/3, -R \leq \mathbf{x}_1 \leq -R/2\}] \\
&\quad\geq \frac{cLR}{24}\left(\|\mathbf{q}^{k\mathbf{w}}\|_2\frac{c}{2} + \|\mathbf{q}^{?\mathbf{w}}\|_2\frac{R}{2}\right)\ ,
\end{aligned}$$

where we used that the distribution is (L, R)-well-behaved. For the case that $c = 1/\|\mathbf{w}\|_2$, we have that

$$I_2 \geq \frac{\tau^2 LR}{72\|\mathbf{w}\|_2}\left(\|\mathbf{q}^{k_{\mathbf{w}}}\|_2 \frac{1}{2\|\mathbf{w}\|_2} + \|\mathbf{q}^{?_{\mathbf{w}}}\|_2 \frac{R}{2}\right) \geq \frac{\tau^2 LR^2}{72\|\mathbf{w}\|_2^{1/2}}\|\mathbf{q}\|_{\mathbf{w}} \,.$$

Finally, for the case that $c = R$, we have that

$$I_2 \geq \frac{\tau^2 LR^2}{72}\left(\|\mathbf{q}^{k_{\mathbf{w}}}\|_2 \frac{R}{2} + \|\mathbf{q}^{?_{\mathbf{w}}}\|_2 \frac{R}{2}\right) \geq \frac{\tau^2 LR^3}{144}\|\mathbf{q}\|_2 \,,$$

where we used that $(a+b)^2 \geq a^2 + b^2$ for $a, b \geq 0$. This completes the proof of Lemma 40. ■

To prove Proposition 15 for this case (the case for which the assumptions of Lemma 40 hold), we use the bound of the contribution of the noise to the gradient from Lemma 34, i.e., that $I_1 \geq -\sqrt{8\epsilon/\mu^3}(\xi/L^2)\|\mathbf{w} \ - \mathbf{w}\|_{\mathbf{w}}$, given that $\|\mathbf{w}\|_2 \geq 2/R$. Putting together our estimates for I_1 and I_2, we obtain:

$$\begin{aligned} I_1 + I_2 &\geq \frac{\tau^2 LR^2}{72\|\mathbf{w}\|_2^{1/2}}\|\mathbf{q}\|_{\mathbf{w}}\left(1 - \sqrt{\epsilon\|\mathbf{w}\|_2}\frac{\xi 72\sqrt{8}}{L^3R^2\mu^{3/2}\tau^2}\right) \\ &\geq \frac{\tau^2 LR^2}{144\|\mathbf{w}\|_2^{1/2}}\|\mathbf{q}\|_{\mathbf{w}} \,, \end{aligned} \tag{6}$$

where, we used that $\sqrt{\epsilon\|\mathbf{w}\|_2}\xi^2/(L^3R^2\mu^{3/2}\tau^2)$ is less than a sufficiently small constant from the assumptions of Proposition 15. Observe that from the assumptions of Proposition 15, there exists a constant $C > 0$,such that $\|\mathbf{w}\|_2 \leq CL^6R^6\mu^3\tau^4/(\epsilon\xi^2)$ hence, we have that

$$I_1 + I_2 \geq \sqrt{\epsilon}\|\mathbf{q}\|_{\mathbf{w}} \,,$$

where we used that $L, R, \mu \leq 1$ and $\xi \geq 1$.

Finally, similar to the previous case, for the case that $\|\mathbf{w}\|_2 \leq 2/R$, it holds

$$I_1 + I_2 \geq \frac{\tau^2 LR^3}{144}\|\mathbf{q}\|_2\left(1 - \sqrt{\epsilon}\frac{144\xi}{LR^3\tau^2}\right) \geq \frac{\tau^2 LR^3}{288}\|\mathbf{q}\|_2 \,,$$

where we used that $(1 - \sqrt{\epsilon}\frac{144\xi}{LR^3\tau^2}) \geq 1/2$.

Case 3: angle of $\mathbf{w}$ and $\mathbf{w}^*$ is greater than $\pi/2$ We now handle the case where $\theta(\mathbf{w}, \mathbf{w}^*) \in (\pi/2, \pi)$. This case is very similar to the second case, i.e., Lemma 40.

Lemma 41 *Let D be an (L, R)-well-behaved distribution. For any vector $\mathbf{w} \in \mathbb{R}^d$ with $\theta(\mathbf{w}^*, \mathbf{w}) \in (\pi/2, \pi)$. We show*

- *If $\|\mathbf{w}\|_2 \geq 2/R$, then $I_2 \geq \frac{\tau^2 LR^2}{16k\mathbf{w}k_2^{1/2}}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$.*
- *Otherwise, we have that $I_2 \geq \frac{\tau^2 LR^4}{16}\|\mathbf{w} - \mathbf{w}^*\|_2$.*

Proof Let $c = \min(1/\|\mathbf{w}\|_2, R)$. We consider the $R/2 \leq \mathbf{x}_1 \leq R$ and $c/2 \leq \mathbf{x}_2 \leq c$ so that $0 \geq \mathbf{w} \ \cdot \mathbf{x}$. Similarly to the previous case, using Claim 37, we get that $I_2 \geq \tau^2 \mathbf{E}_{\mathbf{x} \ D_\mathbf{x}}[(\mathbf{w} \cdot \mathbf{x} - r(\mathbf{w} \ \cdot \mathbf{x}))(\mathbf{w} - \mathbf{w} \ ) \cdot \mathbf{x} \, \mathbf{1}\{0 \leq \mathbf{w} \cdot \mathbf{x} \leq 1\}]$ and for this case it holds that $r(\mathbf{w} \ \cdot \mathbf{x}) \leq 0$. Hence, we have that

$$\begin{aligned}
&\mathbf{E}_{\mathbf{x} \ D_\mathbf{x}}[(\mathbf{w} \cdot \mathbf{x} - r(\mathbf{w} \ \cdot \mathbf{x}))(\mathbf{w} - \mathbf{w} \ ) \cdot \mathbf{x} \, \mathbf{1}\{0 \leq \mathbf{w} \cdot \mathbf{x} \leq 1\}] \\
&\quad \geq \mathbf{E}_{\mathbf{x} \ D_\mathbf{x}}[(\mathbf{w} \cdot \mathbf{x} - r(\mathbf{w} \ \cdot \mathbf{x}))(\mathbf{w} - \mathbf{w} \ ) \cdot \mathbf{x} \, \mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq c, R/2 \leq \mathbf{x}_1 \leq R\}] \\
&\quad \geq \frac{c\|\mathbf{w}\|_2}{2} \mathbf{E}_{\mathbf{x} \ D_\mathbf{x}}[(\mathbf{w} - \mathbf{w} \ ) \cdot \mathbf{x} \, \mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq c, R/2 \leq \mathbf{x}_1 \leq R\}] \ .
\end{aligned}$$

Note that $(\mathbf{w} - \mathbf{w} \ ) \cdot \mathbf{x} = (\|\mathbf{w}\|_2 - \cos\theta\|\mathbf{w} \ \|_2)\mathbf{x}_2 + \sin\theta\|\mathbf{w} \ \|_2\mathbf{x}_1 \geq (\|\mathbf{w}\|_2 - \cos\theta\|\mathbf{w} \ \|_2)c/2 + \sin\theta\|\mathbf{w} \ \|_2 R/2$, when $R/2 \leq \mathbf{x}_1 \leq R$ and $c/2 \leq \mathbf{x}_2 \leq c$. Denote by $\mathbf{q} = \mathbf{w} - \mathbf{w}$ we get that $(\|\mathbf{w}\|_2 - \cos\theta\|\mathbf{w} \ \|_2)c/2 + \sin\theta\|\mathbf{w} \ \|_2 R/2 = \|\mathbf{q}^{k_\mathbf{w}}\|_2 c/2 + \|\mathbf{q}^{?_\mathbf{w}}\|_2 R/2$. Furthermore, using that the distribution D is (L, R)-well-behaved, we have that $\mathbf{E}_{\mathbf{x} \ D_\mathbf{x}}[\mathbf{1}\{c/2 \leq \mathbf{x}_2 \leq c, R/2 \leq \mathbf{x}_1 \leq R\}] \geq L\frac{Rc}{4}$. Hence,

$$I_2 \geq \frac{\tau^2 LRc^2\|\mathbf{w}\|_2}{16}\left(\|\mathbf{q}^{k_\mathbf{w}}\|_2 c + \|\mathbf{q}^{?_\mathbf{w}}\|_2 R\right) \ ,$$

which completes the proof of Lemma 41. ■

Therefore, if $\|\mathbf{w}\|_2 \geq 2/R$, we have $I_1 \geq -\sqrt{8\epsilon/\mu^3}(\xi/L^2)\|\mathbf{w} \ - \mathbf{w}\|_\mathbf{w}$. Putting together our estimates for I_1 and I_2, we obtain:

$$\begin{aligned}
I_1 + I_2 &\geq \frac{\tau^2 LR^2}{16\|\mathbf{w}\|_2^{1/2}}\|\mathbf{w} - \mathbf{w} \ \|_\mathbf{w}\left(1 - \sqrt{\epsilon\|\mathbf{w}\|_2}\frac{\xi 16}{L^3R^2\mu^{3/2}\tau^2}\right) \\
&\geq \frac{\tau^2 LR^2}{32\|\mathbf{w}\|_2^{1/2}}\|\mathbf{w} - \mathbf{w} \ \|_\mathbf{w} \ ,
\end{aligned}$$

where we used that $\sqrt{\epsilon\|\mathbf{w}\|_2}\xi/(L^3R^2\mu^{3/2}\tau^2)$ is less than a sufficiently small constant from the assumptions of Proposition 15. Note that from the assumptions of Proposition 15, there exists a constant $C > 0$ such that $\|\mathbf{w}\|_2 \leq CL^6R^6\mu^3\tau^4/(\epsilon\xi^2)$, and hence it holds

$$I_1 + I_2 \geq \sqrt{\epsilon}\|\mathbf{q}\|_\mathbf{w} \ ,$$

where we used that $L, R, \mu \leq 1$ and $\xi \geq 1$.

Finally, similarly to the previous case, we have that if $\|\mathbf{w}\|_2 \leq 2/R$, it holds that

$$I_1 + I_2 \geq \frac{\tau^2 LR^4}{16}\|\mathbf{w} - \mathbf{w} \ \|_2\left(1 - \sqrt{\epsilon}\frac{16\xi}{LR^4\tau^2}\right) \geq \frac{LR^4\tau^2}{32}\|\mathbf{w} - \mathbf{w} \ \|_2 \ ,$$

where we used that $(1 - \sqrt{\epsilon}\frac{16\xi}{LR^4\tau^2}) \geq 1/2$. This completes the proof of Proposition 15. ■

A.5. Proof of Claim 31

We restate and prove the following claim.

Claim 42 *If $c^{\prime}\kappa/(\epsilon\Lambda_2) \leq \|\mathbf{w}\|_2 \leq c^{\prime}\kappa/\epsilon$ and $\|\mathrm{proj}_{\mathbf{w}^\star}\mathbf{w}\ \|_2 \geq 2K/R$ then*

$$\nabla F_\rho(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}\ ) \geq (c^{\prime}/2)\sqrt{\epsilon}\|\mathbf{w}\ -\mathbf{w}\|_{\mathbf{w}}\,.$$

Proof First, we calculate the contribution of the regularizer in the direction $\mathbf{w}-\mathbf{w}$. This is equal to $2\rho\|\mathbf{w}\|_2^2(\|\mathbf{w}\|_2 - \|\mathbf{w}\ \|_2\cos\theta)$. Note that this is positive when $\|\mathbf{w}\|_2 - \|\mathbf{w}\ \|_2\cos\theta \geq 0$ and negative otherwise. Hence, if $\theta \in [\pi/2,\pi)$ the contribution of the regularizer is positive, and therefore it is bounded below by the contribution of the gradient without the regularizer. Moreover, if $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}\ \|_2$, then $\|\mathbf{w}\|_2 - \|\mathbf{w}\ \|_2\cos\theta \geq \|\mathbf{w}\|_2 - \|\mathbf{w}\ \|_2 \geq 0$, therefore we can again bound from below the contribution of the gradient like before. For the rest of the proof, we consider the case where $\theta \in (0,\pi/2)$ and $\|\mathbf{w}\|_2 \leq 2\|\mathbf{w}\ \|_2$.

From Proposition 15 and specifically Equation (6), we have that as long as $2/R \leq \|\mathbf{w}\|_2 \leq c^{\prime}\kappa/(\epsilon$ and $\|\mathrm{proj}_{\mathbf{w}^\star}\mathbf{w}\ \|_2 \geq 2K/R$ then $\nabla F(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}\ ) \geq c^{\prime}\frac{\tau^2 LR^2}{\kappa\|\mathbf{w}\|_2^{1/2}}\|\mathbf{w}-\mathbf{w}\ \|_{\mathbf{w}}$. Therefore, we have that

$$\begin{aligned}
\nabla F_\rho(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}\ ) &\geq c^{\prime}\tau^2 LR^2\frac{\|\mathbf{w}\ -\mathbf{w}\|_{\mathbf{w}}}{\|\mathbf{w}\|_2^{1/2}} + \rho\|\mathbf{w}\|_2(\|\mathbf{w}\|_2 - \|\mathbf{w}\ \|_2\cos\theta)\\
&\geq c^{\prime}\tau^2 LR^2\frac{\|\mathbf{w}\ -\mathbf{w}\|_{\mathbf{w}}}{\|\mathbf{w}\|_2^{1/2}} - \rho\|\mathbf{w}\|_2\|\mathbf{w}\ \|_2\\
&\geq c^{\prime}\tau^2 LR^2\frac{\|(\mathbf{w}-\mathbf{w}\ )^{\kappa_{\mathbf{w}}}\|_2}{\|\mathbf{w}\|_2^2} + c^{\prime}\tau^2 LR^2\frac{1}{\|\mathbf{w}\|_2}\left(\|\mathrm{proj}_{\mathbf{w}^\star}\mathbf{w}\ \|_2 - \frac{\rho\|\mathbf{w}\|_2^2\|\mathbf{w}\ \|_2}{c^{\prime}\tau^2 LR^2}\right).
\end{aligned}$$

To bound the $\|\mathrm{proj}_{\mathbf{w}^\star}\mathbf{w}\ \|_2 - \rho\|\mathbf{w}\|_2^2\|\mathbf{w}\ \|_2/(c^{\prime}\tau^2 LR^2)$, we have that

$$\frac{\|\mathrm{proj}_{\mathbf{w}^\star}\mathbf{w}\ \|_2 c^{\prime}\tau^2 LR^2}{2\rho\|\mathbf{w}\|_2^2\|\mathbf{w}\ \|_2} \geq \frac{Kc^{\prime}\tau^2 LR}{\rho\|\mathbf{w}\|_2^2\|\mathbf{w}\ \|_2} \geq \frac{KF^3(\mathbf{w}\ )LR\tau^2}{\rho c^{\prime 2}\kappa^2 U} \geq 1\,,$$

where the in the first inequality we used that $\|\mathrm{proj}_{\mathbf{w}^\star}\mathbf{w}\ \|_2 \geq 2K/R$; in the second that $\|\mathbf{w}\|_2 \leq c^{\prime}\kappa/\epsilon$ and $\|\mathbf{w}\ \|_2\epsilon \leq U$; and in the last inequality that $\rho \leq \frac{K\epsilon^3 LR\tau^2}{c^{\prime 2}\kappa^2 U}$. Therefore, we have that $\left(\|\mathrm{proj}_{\mathbf{w}^\star}\mathbf{w}\ \|_2 - 2\rho\|\mathbf{w}\|_2^2\|\mathbf{w}\ \|_2/(c^{\prime}LR^2\tau^2)\right) \geq \|\mathrm{proj}_{\mathbf{w}^\star}\mathbf{w}\ \|_2/2$ and the result follows similar to the proof of Claim 30. ■

A.6. Parameter vs L_2^2 Distance

Lemma 43 (Parameter vs L_2^2 Distance) *Let $D_{\mathbf{x}}$ be an (L,R) well-behaved distribution. Let σ be a (τ,ξ,μ)-sigmoidal activation. For any vectors $\mathbf{w},\mathbf{v}\in\mathbb{R}^d$, we have $\mathbf{E}_{\mathbf{x}\ D_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))^2] \leq \xi^2\|\mathbf{w}-\mathbf{v}\|_2^2$. Moreover, if $\theta=\theta(\mathbf{w},\mathbf{v}) < \pi/4$, $\|\mathbf{w}\|_2 \leq \delta\|\mathbf{v}\|_2$ and $\delta \geq 1$, and $\|\mathbf{w}\|_2 > 2/R$, it holds*

$$\mathbf{E}_{\mathbf{x}\ D_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))^2]\ .\ \frac{\xi^2\delta^3}{L^4\mu^3}\|\mathbf{w}-\mathbf{v}\|_{\mathbf{w}}^2\,.$$

Proof First, we consider the case where $\|\mathbf{w}\|_2 \geq 2/R$ and $\theta \in (0, \pi/4)$. Let V be the subspace spanned by $\mathbf{w}, \mathbf{v}$ and assume without loss of generality that $\mathbf{w}/\|\mathbf{w}\|_2 = \mathbf{e}_2$ and therefore $\mathbf{v} = \|\mathbf{v}\|_2(\cos\theta \mathbf{e}_2 - \sin\theta \mathbf{e}_1)$, we abuse the notation of $D_{\mathbf{x}}$ to be the distribution projected on V. It holds that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{v}\cdot\mathbf{x}))^2] &= \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}\left[\left(\int_{\mathbf{v}\cdot\mathbf{x}}^{\mathbf{w}\cdot\mathbf{x}} \sigma^{\emptyset}(t)\mathrm{d}t\right)^2\right] \\
&\leq \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}\left[\left(\int_{\mathbf{v}\cdot\mathbf{x}}^{\mathbf{w}\cdot\mathbf{x}} \xi\exp(-\mu|t|)\mathrm{d}t\right)^2\right] \\
&\leq \xi^2 \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{v}\cdot\mathbf{x})^2(\exp(-2\mu|\mathbf{w}\cdot\mathbf{x}|) + \exp(-2\mu|\mathbf{v}\cdot\mathbf{x}|))]\,,
\end{aligned}$$

where we used that σ is non-decreasing. Let $\mathbf{q} = \mathbf{w} - \mathbf{v}$. Using Claim 36, it holds that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{v}\cdot\mathbf{x})^2\exp(-2\mu|\mathbf{w}\cdot\mathbf{x})|)] &= \mathbf{E}[((\mathbf{q}^{k_{\mathbf{w}}}\cdot\mathbf{x})^2 + (\mathbf{q}^{?_{\mathbf{w}}}\cdot\mathbf{x})^2)\exp(-2\mathbf{w}\cdot\mathbf{x}\mu)] \\
&\leq \frac{8}{L^4}(\|\mathbf{q}^{k_{\mathbf{w}}}\|_2^2\frac{1}{\mu^3\|\mathbf{w}\|_2^3} + \|\mathbf{q}^{?_{\mathbf{w}}}\|_2^2\frac{1}{\mu\|\mathbf{w}\|_2}) \leq \frac{8\|\mathbf{q}\|_{\mathbf{w}}^2}{L^4\mu^3}\,.
\end{aligned}$$

Moreover, for the second term, it holds that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{v}\cdot\mathbf{x})^2\exp(-2\mu|\mathbf{v}\cdot\mathbf{x})|)] &= \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{q}^{k_{\mathbf{w}}}\cdot\mathbf{x})^2 + (\mathbf{q}^{?_{\mathbf{w}}}\cdot\mathbf{x})^2)\exp(-2\mathbf{v}\cdot\mathbf{x}\mu)] \\
&\leq \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{q}^{k_{\mathbf{w}}}\cdot\mathbf{x})^2 + (\mathbf{q}^{?_{\mathbf{w}}}\cdot\mathbf{x})^2)\exp(-2\|\mathbf{v}\|_2\cos\theta\mathbf{x}_2\mu)] \\
&\leq \frac{8}{L^4}(\|\mathbf{q}^{k_{\mathbf{w}}}\|_2^2\frac{1}{\mu^3\|\mathbf{v}\|_2^3\cos\theta^3} + \|\mathbf{q}^{?_{\mathbf{w}}}\|_2^2\frac{1}{\mu\|\mathbf{v}\|_2\cos\theta})\ \cdot\ \frac{\delta^3\|\mathbf{q}\|_{\mathbf{w}}^2}{L^4\mu^3}\,,
\end{aligned}$$

where in the last inequality, we used that $\cos\theta \geq 1/2$ and that $\|\mathbf{w}\|_2 \leq \delta\|\mathbf{v}\|_2$. For the other cases, we use the "trivial" upper-bound, i.e.,

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{v}\cdot\mathbf{x}))^2] &\leq \xi^2 \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{v}\cdot\mathbf{x})^2] \\
&\leq \xi^2\|\mathbf{w} - \mathbf{v}\|_2^2 \sup_{k\mathbf{w}k_2=1} \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x})^2] \\
&\leq \xi^2\|\mathbf{w} - \mathbf{v}\|_2^2\,,
\end{aligned}$$

where we used that σ is ξ-Lipschitz and that the distribution $D_{\mathbf{x}}$ is isotropic. ■

Appendix B. Omitted Proofs of Section 4

B.1. Proof of Proposition 21

We restate and prove the following claim.

Proposition 44 *Let D be an (ϵ, W)-corrupted, (L, R)-well-behaved distribution and σ be an (α, λ)-unbounded activation. For any $\mathbf{w} \in \mathbb{R}^d$ with $\mathbf{w}\cdot\mathbf{w}\ \ \geq 0$ and $\|\mathbf{w} - \mathbf{w}\ \|_2 \geq C\lambda/(\alpha^2 LR^4)\sqrt{\epsilon}$, it holds $\nabla F(\mathbf{w})\cdot(\mathbf{w} - \mathbf{w}\ )\ \&\ \ \alpha^2 LR^4\|\mathbf{w} - \mathbf{w}\ \|_2^2$.*

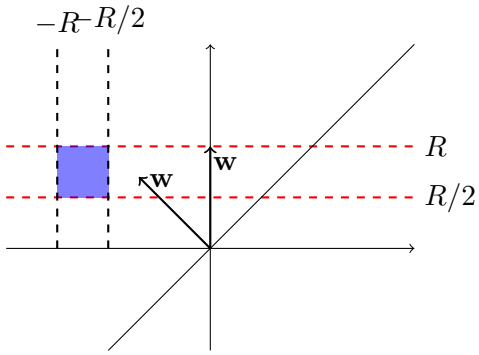


Figure 4: Using our distributional assumptions, we identify a region ("blue") that provides enough contribution to the gradient, so that an update step will decrease the distance with the optimal one.

Proof We have

$$\begin{aligned}\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) &= \underbrace{\mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})]}_{I_1} \\ &+ \underbrace{\mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - \sigma(\mathbf{w}^* \cdot \mathbf{x}))\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})]}_{I_2} .\end{aligned}$$

We start by bounding the contribution of I_1. We show that the "noisy" integral I_1 has small negative contribution that is bounded by some multiple of $\|\mathbf{w} - \mathbf{w}^*\|_2\sqrt{\epsilon}$. We show the following claim.

Claim 45 *It holds that* $I_1 \geq -2\lambda\|\mathbf{w} - \mathbf{w}^*\|_2\sqrt{\epsilon}$.

Proof Using the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}I_1 &\geq -\mathbf{E}[|(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})|] \\ &\geq -\lambda\,\mathbf{E}[|\sigma(\mathbf{w} \cdot \mathbf{x}) - y||\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x}|] \\ &\geq -\lambda(\mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2])^{1/2}(\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2])^{1/2} \\ &\geq -\lambda\sqrt{\epsilon}\,\|\mathbf{w} - \mathbf{w}^*\|_2 \left(\max_{\mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\|_2 = 1} \mathbf{E}[(\mathbf{v} \cdot \mathbf{x})^2]\right)^{1/2} \geq -2\lambda\|\mathbf{w} - \mathbf{w}^*\|_2\sqrt{\epsilon} ,\end{aligned}$$

where in the first inequality we used the fact that σ is λ-Lipschitz, in the second that $F(\mathbf{w}^*) = \epsilon$ and in the last the fact that the distribution is isotropic. Hence, we have that

$$I_1 \geq -2\lambda\|\mathbf{w} - \mathbf{w}^*\|_2\sqrt{\epsilon} . \tag{7}$$

This completes the proof of Claim 45. ■

Next we bound from below the contribution of the "clean" examples, i.e., I_2. We show that this positive contribution is bounded below by a multiple of $\|\mathbf{w} - \mathbf{w}^*\|_2^2$, which is enough to surpass the contribution of the negative region I_2. We have the following claim.

Claim 46 *It holds that* $I_2 \geq \frac{7\alpha^2 LR^4}{64}\|\mathbf{w} - \mathbf{w}^*\|_2^2$.

Proof We first notice that $(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}) \geq 0$ for every $\mathbf{x}\in\mathbb{R}^d$, because σ is an increasing function. Without loss of generality, we may assume that $\mathbf{w}/\|\mathbf{w}\|_2 = \mathbf{e}_2$ and therefore $\mathbf{w}^* = \|\mathbf{w}^*\|_2(\cos\theta\mathbf{e}_2 - \sin\theta\mathbf{e}_1)$. For simplicity, we abuse notation and denote $D_{\mathbf{x}}$ the marginal distribution on the subspace spanned by the vectors $\mathbf{w}, \mathbf{w}^*$. From the definition (α,λ)-unbounded activations, we have that $\inf_{t\in(0,R/\|\mathbf{w}\|_2)}\sigma'(t) \geq \inf_{t\in(0,1)}\sigma'(t) \geq \alpha > 0$. We now have that

$$\begin{aligned} I_2 &= \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})] \\ &\geq \alpha^2\,\mathbf{E}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbf{1}\{\mathbf{x}_1\in(-R,-R/2), \mathbf{x}_2\in(R/2,R)\}]\,, \end{aligned}$$

where the region inside the indicator is the "blue region" of Figure 4. We remark that this region is not particularly "special": we can use other regions that contain enough mass depending on the distributional assumptions. At a high level, we only need a rectangle that contains enough mass and the gradient in these points is non-zero. Therefore, using the fact that the distribution is (L,R)-well-behaved, we have

$$\begin{aligned} &\mathbf{E}\left[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\,\mathbf{1}\{\mathbf{x}_1\in(-R,-R/2), \mathbf{x}_2\in(R/2,R)\right] \\ &\quad= \mathbf{E}\left[(\|\mathbf{q}^{k_{\mathbf{w}}}\|_2^2\mathbf{x}_2^2 + \|\mathbf{q}^{?_{\mathbf{w}}}\|_2^2\mathbf{x}_1^2)\,\mathbf{1}\{\mathbf{x}_1\in(-R,-R/2), \mathbf{x}_2\in(R/2,R)\}\right] \\ &\quad\geq \frac{7LR^4}{32}(\|\mathbf{q}^{k_{\mathbf{w}}}\|_2^2 + \|\mathbf{q}^{?_{\mathbf{w}}}\|_2^2) = \frac{7LR^4}{64}\|\mathbf{w} - \mathbf{w}^*\|_2^2\,, \end{aligned}$$

where $\mathbf{q} = \mathbf{w} - \mathbf{w}^*$. Therefore, we proved that

$$I_2 \geq \frac{7\alpha^2 LR^4}{64}\|\mathbf{w} - \mathbf{w}^*\|_2^2\,. \tag{8}$$

■

Thus, combining Equations (7) and (8), we have that

$$\nabla F(\mathbf{w})\cdot(\mathbf{w} - \mathbf{w}^*) \geq \|\mathbf{w} - \mathbf{w}^*\|_2\left(\alpha^2\frac{7LR^4}{64}\|\mathbf{w} - \mathbf{w}^*\|_2 - 2\lambda\sqrt{\epsilon}\right)\,,$$

which completes the proof of Proposition 21. ■

B.2. Proof of Theorem 19

We restate and prove the following theorem:

Theorem 47 (Stationary Points of (α,λ)-Unbounded Activations) *Let D be an (ϵ, W)-corrupted, (L,R)-well-behaved distribution in $\mathbb{R}^d$. Let σ be an (α,λ)-unbounded activation and let $F(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim D}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2]$. Then, if for some $\mathbf{w}\in\mathbb{R}^d$, with $\mathbf{w}\cdot\mathbf{w}^* \geq 0$ and $\|\mathbf{w}\|_2 \leq W$ it holds $\|\nabla F(\mathbf{w})\|_2 \leq 2\lambda\sqrt{\epsilon}$, then $F(\mathbf{w}) \leq C\left(\frac{\lambda}{\alpha}\right)^4\frac{1}{L^2R^8}\epsilon$, for some absolute constant $C > 0$.*

Remark 48 We remark that we assume $\sigma(0) = 0$ to simplify the presentation. If $\sigma(0) \neq 0$, then we can always consider the loss function $\tilde{\sigma}(t) = \sigma(t) - \sigma(0)$ and also subtract $\sigma(0)$ from the labels y, i.e., $y' = y - \sigma(0)$. Our results directly apply to the transformed instance.

To prove Theorem 19, in order to reach a contradiction we assume that $F(\mathbf{w}) \geq C\frac{\lambda^4}{\alpha^4 L^2 R^8}\epsilon$ for absolute constant $C > 0$ and that $\|\nabla F(\mathbf{w})\|_2 \leq \frac{\alpha^2 L R^4}{100}\sqrt{\epsilon}$. From Proposition 21, we have that

$$\frac{7\alpha^2 L R^4}{64}\sqrt{\epsilon} \geq \nabla F(\mathbf{w}) \cdot \frac{(\mathbf{w} - \mathbf{w}^*)}{\|\mathbf{w} - \mathbf{w}^*\|_2} \geq \frac{7\alpha^2 L R^4}{64}\left(\|\mathbf{w} - \mathbf{w}^*\|_2 - \sqrt{\epsilon}\frac{\lambda 128}{7\alpha^2 L R^4}\right) ,$$

and therefore

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \frac{128\lambda\sqrt{\epsilon}}{7\alpha^2 L R^4} . \tag{9}$$

Moreover, we have

$$\begin{aligned} F(\mathbf{w}) = \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] &\leq 2\,\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] + 2\,\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] \\ &\leq 2\epsilon + 2\lambda^2\|\mathbf{w} - \mathbf{w}^*\|_2^2 . \end{aligned}$$

Therefore, using Equation (9), we get

$$F(\mathbf{w}) \leq C\left(\frac{\lambda}{\alpha}\right)^4 \frac{1}{L^2 R^8}\epsilon ,$$

which leads to a contradiction. This completes the proof of Theorem 19.

Appendix C. Omitted Proofs of Section 5

C.1. Proof of Theorem 22

We restate and prove the following theorem:

Theorem 49 (Learning Sigmoidal Activations) *Let D be an ϵ-corrupted, (L, R)-well-behaved distribution on $\mathbb{R}^d \times \mathbb{R}$ and $\sigma(\cdot)$ be a (τ, μ, ξ)-sigmoidal activation. Set $\kappa = L^6 R^6 \mu^3 \tau^4/\xi^2$ and let $c > 0$ be a sufficiently small absolute constant. Then gradient descent (Algorithm 1) with step size $\eta = c\epsilon^{2.5}$, regularization $\rho = (1/c)\epsilon^3/k^5$, truncation threshold $M = \xi/\mu$, $N = \widetilde{\Theta}(d/\epsilon \log(1/\delta))\,\mathrm{poly}(1/\kappa)$ samples, and $T = \mathrm{poly}(1/(\epsilon\kappa))$ iterations converges to a vector $\mathbf{w}^{(T)} \in \mathbb{R}^d$ that, with probability $1 - \delta$, satisfies $F(\mathbf{w}^{(T)}) \leq \mathrm{poly}(1/\kappa)\,\epsilon$.*

We will first assume that we have access to the population gradients of the L_2^2 objective and then show that given samples from D, gradient descent again converges to some approximately optimal solution. Before we start, we observe that, without loss of generality, we may assume that $\sigma(0) = 0$ (otherwise we can subtract $\sigma(0)$ from y and the activation, see Remark 48). Moreover, from Fact 28 we know that $|\sigma(t)| \leq \xi/\mu$. Therefore, for every example $(\mathbf{x}, y) \sim D$ we can "truncate" its label to $\widehat{y} = \mathrm{sign}(y)\min(|y|, \xi/\mu)$ and the new instance will still be (at most) ϵ-corrupted. To simplify notation, from now on, we will overload the notation and use y instead of $\widehat{y}$ (assuming that for every example $(\mathbf{x}, y)$ it holds that $|\sigma(\mathbf{w}\cdot\mathbf{x}) - y| \leq 2\xi/\mu$).

Population Gradient Descent Recall that in Proposition 14 we showed that the population gradient field "points" to the right direction, i.e., a step of gradient descent would decrease the distance between the current $\mathbf{w}$ and a target $\mathbf{v}$. Notice that in Proposition 14 we require that the target vector $\mathbf{v}$ satisfies $F(\mathbf{v}) = O(\epsilon)$ and $\|\mathbf{v}\|_2 \leq O(1/\epsilon)$. Indeed, from Lemma 13, we have that there exists an approximately optimal target vector $\mathbf{v} \in \mathbb{R}^d$, such that $\|\mathbf{v}\|_2 \leq U/\epsilon$ for some $U = O(\xi/(\mu L))$ and $F(\mathbf{v}) = O(\epsilon)$.

Moreover, recall the way we measure distances changes depending on how far $\mathbf{w}$ is from the origin (see the three cases of Proposition 14). We first show that, given any gradient field that satisfies the properties given in Proposition 14, then gradient descent with an appropriate step size will converge to the target vector $\mathbf{v}$. The distance of $\mathbf{v}$ and the guess $\mathbf{w}^{(T)}$ after T gradient iterations will be measured with respect to $\|\cdot\|_2$ when $\mathbf{w}^{(T)}$ is close to the origin and with respect to $\|\cdot\|_{\mathbf{w}^{(T)}}$ (see Definition 10) otherwise.

Lemma 50 (Gradient Field Distance Reduction) *Let $Z_1 > Z_0 \geq 1$. Let $\mathbf{g} : \mathbb{R}^d \mapsto \mathbb{R}^d$ be a vector field with $\|\mathbf{g}(\mathbf{w})\|_2 \leq B$ for every $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq 2Z_1$. We assume that $\mathbf{g}$ satisfies the following properties with respect to some unknown target vector $\mathbf{v}$:*

1. *If $\|\mathbf{w}\|_2 \leq Z_0$ and $\|\mathbf{w} - \mathbf{v}\|_2 \geq \alpha_1$ then it holds that $\mathbf{g}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) \geq \alpha_2 \|\mathbf{w} - \mathbf{v}\|_2$, for $\alpha_1, \alpha_2 > 0$.*
2. *If $Z_0 < \|\mathbf{w}\|_2 \leq Z_1$ and $\|\mathbf{w} - \mathbf{v}\|_{\mathbf{w}} \geq \beta_1$ it holds that $\mathbf{g}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) \geq \beta_2 \|\mathbf{w} - \mathbf{v}\|_{\mathbf{w}}$, for $\beta_1, \beta_2 > 0$.*
3. *For some $\zeta \in (0,1), \gamma > 0$, we have that if $\|\mathbf{w}\|_2 \geq \zeta Z_1$, it holds that $\mathbf{g}(\mathbf{w}) \cdot \mathbf{w} \geq \gamma \|\mathbf{w}\|_{\mathbf{w}}$.*

We consider the update rule $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \mathbf{g}(\mathbf{w}^{(t)})$ initialized with $\mathbf{w}^{(0)} = \mathbf{0}$ and step size

$$\eta = \frac{1}{B^2} \min\left(\alpha_1\alpha_2, \frac{\beta_1\beta_2}{Z_1^{3/2}}, \frac{2\gamma}{Z_1}, (1-\zeta)Z_1 B\right) .$$

Let T be any integer larger than $\left\lceil \frac{\|\mathbf{v}\|^2}{\eta \min(\alpha_1\alpha_2, \beta_1\beta_2/Z_1^{3/2})} \right\rceil$. We have that $\|\mathbf{w}^{(T)}\|_2 \leq Z_1$. Moreover, if $\|\mathbf{w}^{(T)}\|_2 \leq Z_0$ it holds that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq \eta B + \max(\alpha_1, (2Z_0)^{3/2}\beta_1)$ and if $\|\mathbf{w}^{(T)}\|_2 > Z_0$ we have $\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \sqrt{2}\eta B + \max(\sqrt{2}\alpha_1, e^{3Z_1^{3/2}\eta B}\beta_1)$.

We now show that the population gradient field, $\nabla F(\mathbf{w})$ satisfies the assumptions of Lemma 54 for any $\|\mathbf{w}\|_2 \leq \kappa/\epsilon$. We first show that $\|\nabla F(\mathbf{w})\|_2$ is bounded. We have

$$\begin{aligned}
\|\nabla F(\mathbf{w})\|_2 &= \left\|\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}] + \rho\|\mathbf{w}\|_2\mathbf{w}\right\|_2 \\
&\leq \max_{\|\mathbf{u}\|_2=1} \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] + \rho\|\mathbf{w}\|_2^2 \\
&\leq (\xi^2/\mu) \max_{\|\mathbf{u}\|_2=1} \mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|] + \rho\|\mathbf{w}\|_2^2 \leq \xi^2/\mu + \rho\|\mathbf{w}\|_2^2 \;.\; 1/\kappa ,
\end{aligned}$$

where we used the fact that $|\sigma(\mathbf{w}\cdot\mathbf{x}) - y| \leq \xi/\mu$, $\sigma'(t) \leq \xi$, that the $\mathbf{x}$-marginal of D is isotropic and that $\rho = O(\epsilon^3)$.

From Proposition 14 we have that for any vector $\mathbf{w} \in \mathbb{R}^d$, with $\|\mathbf{w}\|_2 \leq 2/R$, it holds that if $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq \sqrt{\epsilon}/(c'\kappa)$ for some absolute constant $c' > 0$ it holds $\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq \sqrt{\epsilon}\|\mathbf{w} -$

$\mathbf{w}\ \|_2$. Furthermore, for any vector $\mathbf{w} \in \mathbb{R}^d$, with $2/R \leq \|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$, if $\|\mathbf{w}-\mathbf{w}\ \|_{\mathbf{w}} \geq c'U\sqrt{\epsilon}/\kappa$, then $\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w}-\mathbf{w}\ ) \geq c'\sqrt{\epsilon}\|\mathbf{w}-\mathbf{w}\ \|_{\mathbf{w}}$. Finally, for any vector $\mathbf{w} \in \mathbb{R}^d$, with $c''\kappa/\epsilon \leq \|\mathbf{w}\|_2$, it holds that $\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} \geq c'\sqrt{\epsilon}\|\mathbf{w}\|_{\mathbf{w}}$, for some $0 < c'' \leq c'$. Therefore, the true gradient field of F satisfies the assumptions of Lemma 54 with $B = O(1/\kappa)$, $Z_0 = O(1/R)$, $Z_1 = \mathrm{poly}(LR\tau\mu/\xi)$, $\zeta = 1/2$, $\alpha_1, \beta_1 = \mathrm{poly}(\xi/(\mu\tau LR))\sqrt{\epsilon}$, and $\alpha_2, \beta_2, \gamma = O(\sqrt{\epsilon})$.

Using Sample-Estimated Gradients In the following claim we show that since $\nabla F(\mathbf{w})$ is a sub-exponential random variable, with roughly $N = \widetilde{O}(d/\epsilon'^2)$ samples we can get ϵ-estimates of the population gradients. The proof of Claim 51 can be found in Appendix C.

Claim 51 (Sample Complexity of Gradient Estimation) *Fix $B, \epsilon' > 0$ with $\epsilon' \leq 1/\sqrt{B}$. Using $N = \widetilde{O}((d/\epsilon'^2)\mathrm{poly}(\xi/(L\mu))\log(B/\delta))$ samples from D, we define the empirical L_2^2-loss as $\widehat{F}(\mathbf{w}) = \sum_{i=1}^N (\sigma(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})^2$ and its corresponding empirical gradient field $\nabla\widehat{F}(\mathbf{w})$. With probability at least $1-\delta$, for all $\mathbf{w}$ with $\|\mathbf{w}\| \leq B$, it holds $\|\nabla\widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \epsilon'$ and $\|\nabla\widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|\ _{,\mathbf{w}} \leq \epsilon'$.*

Using Claim 51, we can estimate all the gradients with probability $1-\delta'$ and accuracy ϵ', with $N = \widetilde{O}((d/\epsilon'^2)\mathrm{poly}(\xi/(L\mu))\log(B/\delta'))$ samples, for some parameters $\epsilon', \delta > 0$ that we will choose below. We now show that the empirical gradients will also satisfy the required gradient field assumptions of Lemma 54. Assume that we have $\|\nabla\widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \epsilon'$ and $\|\nabla\widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|\ _{,\mathbf{w}} \leq \epsilon'$ for some ϵ' to be specified later. Using the triangle inequality we have that

$$\|\nabla\widehat{F}(\mathbf{w})\|_2 \leq \|\nabla F(\mathbf{w})\|_2 + \epsilon' .$$

Therefore, for $\epsilon' \leq 1/\kappa$ we obtain that $\|\nabla\widehat{F}(\mathbf{w})\|_2 = O(1/\kappa) \leq O(B)$. We next show that the empirical gradient also points to the direction of $\mathbf{w}-\mathbf{w}\ $, i.e., satisfies the assumptions of Lemma 50. For the case where $\|\mathbf{w}\| \leq Z_0$ we have that

$$\nabla\widehat{F}(\mathbf{w}) \cdot (\mathbf{w}-\mathbf{v}) \geq \nabla F(\mathbf{w}) \cdot (\mathbf{w}-\mathbf{v}) - \epsilon'\|\mathbf{w}-\mathbf{v}\|_2 \geq (\alpha_2 - \epsilon')\|\mathbf{w}-\mathbf{v}\|_2 . \tag{10}$$

Therefore, we need to choose $\epsilon' < \alpha_2$. Similarly, for the case where $Z_0 < \|\mathbf{w}\| \leq Z_1$ we have that

$$\begin{aligned}\nabla\widehat{F}(\mathbf{w}) \cdot (\mathbf{w}-\mathbf{v}) &\geq \nabla F(\mathbf{w}) \cdot (\mathbf{w}-\mathbf{v}) - (\nabla\widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})) \cdot (\mathbf{w}-\mathbf{v}) \\ &\geq \beta_2\|\mathbf{w}-\mathbf{v}\|_{\mathbf{w}} - \|\nabla\widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|\ _{,\mathbf{w}}\|\mathbf{w}-\mathbf{v}\|_{\mathbf{w}} \\ &\geq (\beta_2 - \epsilon')\|\mathbf{w}-\mathbf{v}\|_{\mathbf{w}} .\end{aligned} \tag{11}$$

Finally, for $\|\mathbf{w}\|_2 \geq \zeta Z_1$ we similarly get the lower bound

$$\nabla\widehat{F}(\mathbf{w}) \cdot \mathbf{w} \geq (\gamma - \epsilon')\|\mathbf{w}\|_{\mathbf{w}} . \tag{12}$$

Therefore, it suffices to choose $\epsilon' \leq \min(\alpha_2, \beta_2, \gamma)/2 = (\sqrt{\epsilon})\mathrm{poly}(LR\mu\tau/\xi)$. Assuming that all the empirical gradients used by the gradient descent satisfy the error bound with ϵ' as above, from Lemma 50, we obtain that with step size $\eta = \mathrm{poly}(\epsilon\tau\mu LR/\xi)$ after $T = \mathrm{poly}(1/(\epsilon\eta))$ iterations, we will have that if $\mathbf{w}^{(T)} \leq Z_0$, then $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq \mathrm{poly}(\xi/(LR\tau\mu))\sqrt{\epsilon}$ which using Lemma 43, implies that $F(\mathbf{w}^{(T)}) \leq \mathrm{poly}(\xi/(LR\mu\tau)(\epsilon)$. Similarly, if $\|\mathbf{w}^{(T)}\|_2 > Z_0$, we obtain that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \mathrm{poly}(\xi/(LR\tau\mu))\sqrt{\epsilon}$ which again implies that $F(\mathbf{w}^{(T)}) \leq \mathrm{poly}(\xi/(LR\mu\tau))\epsilon$.

C.2. Proof of Theorem 23

We restate and prove the following claim.

Theorem 52 (Unbounded Activations) *Let D be an (ϵ, W)-corrupted, (L, R)-well-behaved distribution on $\mathbb{R}^d \times \mathbb{R}$ and $\sigma(\cdot)$ be an (α, λ)-unbounded activation. Set $\kappa = \mathrm{poly}(LR\alpha/\lambda)/(W^2 \log(W))$. The gradient descent Algorithm 2 with step size $\eta = \kappa\epsilon$, truncation threshold $M = \widetilde{O}((W/L) \max(\log(\lambda^2 W^2/\epsilon), 1))$, $N = \widetilde{\Theta}(d/\kappa \log(1/\delta) \max(\mathrm{poly}\log(1/\epsilon), 1))$ samples, and $T = \mathrm{poly}(\log(1/\epsilon), 1/\kappa)$ iterations converges to a vector $\mathbf{w}^{(T)} \in \mathbb{R}^d$ that, with probability $1 - \delta$, satisfies $F(\mathbf{w}^{(T)}) \; . \; \frac{1}{LR^4}\left(\frac{\lambda}{\alpha}\right)^4 \epsilon$.*

Proof We consider the population loss $F(\mathbf{w}) = (1/2)\, \mathbf{E}_{(\mathbf{x},y) \;\; D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. As we discussed in Remark 11, we can assume that $\lambda \geq 1$ and $L \leq 1$ for making the presentation simpler. Moreover, we can also assume that $W \geq 1$, because if $W < 1$, then $\inf_{\|\mathbf{w}\|_2 \;\; 1} F(\mathbf{w}) \leq \inf_{\|\mathbf{w}\|_2 \;\; W} F(\mathbf{w})$, therefore the distribution is $(\epsilon, 1)$-corrupted. Denote $\mathbf{w} \;\;$ to be a vector with $\|\mathbf{w} \;\|_2 \leq W$ that achieves $F(\mathbf{w} \;) \leq \epsilon$. First, we assume that $\epsilon \geq \lambda^2 W^2/C$, for some $C \geq 10$. Then, any solution $\mathbf{w} \in \mathbb{R}^d$, gets error $F(\mathbf{w}) \leq 2\epsilon$. To see this, note that

$$F(\mathbf{w}) \leq 2F(\mathbf{w} \;) + \mathop{\mathbf{E}}_{\mathbf{x} \;\; D_{\mathbf{x}}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - \sigma(\mathbf{w} \;\; \cdot \mathbf{x}))^2] \leq 2\epsilon + 4\lambda^2 W^2 \leq 2C\epsilon \; .$$

Next we consider the case where $\epsilon \leq \lambda^2 W^2/C$. First, using the exponential concentration properties of well-behaved distributions, we observe that we can truncate the labels y such that $|y| \leq M$, for some $M > 0$, without increasing the level of corruption by a lot. Given the exponential concentration of the distribution and the fact that $\|\mathbf{w} \;\|_2 \leq W$, we show that we can pick $M = \Theta(W \max(\log(W/\epsilon), 1))$ so that the instance $\widetilde{D}$ of $(\mathbf{x}, \widetilde{y})$, where $\widetilde{y} = \mathrm{sign}(y) \min(|y|, M)$, is at most $O(\epsilon)$-corrupted. Formally, we show

Claim 53 *Let $M = \Theta(W \max(\log(W/\epsilon), 1))$. Denote by $\mathrm{tr}(y) = \mathrm{sign}(y) \min(|y|, M)$. Then, it holds that:*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y) \;\; D}[(\mathrm{tr}(y) - \sigma(\mathbf{w} \;\; \cdot \mathbf{x}))^2] \leq O(\epsilon) \; .$$

Thus, from now, on we will assume that $|y| \leq M$ and keep denoting the instance distribution as D.

We first show that, given any gradient field that satisfies the properties given in Proposition 21, then gradient descent with an appropriate step size will converge to the target vector $\mathbf{v}$.

Lemma 54 (Gradient Field Distance Reduction) *Let $\mathbf{g} : \mathbb{R}^d \mapsto \mathbb{R}^d$ be a vector field and fix $W, B \geq 1$. We assume that $\mathbf{g}$ satisfies the following properties with respect to some unknown target vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 \leq W$. Fix parameters $\alpha_1, \alpha_2 > 0$. For every $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq 2W$ it holds that $\|g(\mathbf{w})\|_2 \leq B \max(\|\mathbf{w} - \mathbf{v}\|_2, \alpha_1)$. Moreover, if $\|\mathbf{w} - \mathbf{v}\|_2 \geq \alpha_1$ and $\theta(\mathbf{w}, \mathbf{v}) \in (0, \pi/2)$ then it holds that $\mathbf{g}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) \geq \alpha_2 \|\mathbf{w} - \mathbf{v}\|_2^2$. We consider the update rule $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \mathbf{g}(\mathbf{w}^{(t)})$ initialized with $\mathbf{w}^{(0)} = \mathbf{0}$ and step size $\eta \leq \alpha_2/B$. Let T be any integer larger than $\lceil (W^2 + \log(1/\alpha_1))/(\eta\alpha_2) \rceil$, it holds that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq (1 + \eta B)\alpha_1$.*

We first show that Algorithm 2 with the population gradients would converge after $\mathrm{polylog}(1/\epsilon)$ iterations to an approximately optimal solution. We have

$$\|\nabla F(\mathbf{w})\|_2 = \|\, \mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})\mathbf{x}]\|_2 = \max_{\|\mathbf{u}\|_2 = 1} \mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})\mathbf{u} \cdot \mathbf{x}] \; .$$

Let $\mathbf{w}$ be any vector with $\|\mathbf{w}\|_2 \leq W$ and let $\mathbf{u}$ be any unit vector. By adding and subtracting $\sigma(\mathbf{w} \cdot \mathbf{x})$, we get:

$$\begin{aligned}
&\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] \\
&\quad = \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] + \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] \\
&\quad \leq \lambda\,\mathbf{E}[|(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)||\mathbf{u}\cdot\mathbf{x}|] + \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] \\
&\quad \leq \lambda\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2]^{1/2}\,\mathbf{E}[(\mathbf{u}\cdot\mathbf{x})^2]^{1/2} + \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2\sigma'(\mathbf{w}\cdot\mathbf{x})^2]^{1/2}\,\mathbf{E}[(\mathbf{u}\cdot\mathbf{x})^2]^{1/2} \\
&\quad \leq \lambda\sqrt{\epsilon} + \lambda\|\mathbf{w} - \mathbf{w}^*\|_2\,,
\end{aligned}$$

where for the third inequality we used the fact the $\mathbf{x}$-marginal of D is isotropic and that $\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2]^{1/2} \leq \sqrt{\epsilon}$. Hence, we have that $\|\nabla F(\mathbf{w})\|_2 \leq \lambda(\sqrt{\epsilon} + \|\mathbf{w} - \mathbf{w}^*\|_2)$. Using Proposition 21, we have that, the true gradients point in the right direction: we have that for any vector $\mathbf{w} \in \mathbb{R}^d$, with $\theta(\mathbf{w}, \mathbf{w}^*) \in (0, \pi/2)$, it holds that if $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq c'\lambda\sqrt{\epsilon}/(LR^4\alpha^2)$, for some absolute constant $c' > 0$, then it holds that $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq (1/2)\alpha^2 LR^4\|\mathbf{w} - \mathbf{w}^*\|_2^2$. Therefore, the true gradient field of F satisfies the assumptions of Lemma 54 with $B = O(\lambda)$, $\alpha_1 = c'\sqrt{\epsilon}/(LR^4)\lambda/\alpha^2$ and $\alpha_2 = (1/2)\alpha^2 LR^4$.

In the following claim, we show that with roughly $\widetilde{O}(d\|\mathbf{w} - \mathbf{w}^*\|_2^2/\epsilon^2)$ samples in each iteration, we can get ϵ'-estimates of the population gradients.

Claim 55 *Let $\mathbf{w}$ be a vector with $\|\mathbf{w}-\mathbf{w}^*\| \geq \sqrt{\epsilon}$. Using $N = \widetilde{O}(d\lambda^4\|\mathbf{w}-\mathbf{w}^*\|_2^2 W^2 \max(\log^2(1/\epsilon), 1)/(L\epsilon')^2)$ samples, we can compute an empirical gradient $\mathbf{g}(\mathbf{w})$ such that $\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \epsilon$ with probability $1 - \delta$.*

We now show that the empirical gradients will also satisfy the required gradient field assumptions of Lemma 54. Assume that we have $\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \epsilon'$ for some ϵ' to be specified later. Using the triangle inequality we get for any $\mathbf{w}$ with $\|\mathbf{w}\| \leq 2W$ it holds that

$$\|\mathbf{g}(\mathbf{w})\|_2 \leq \|\nabla F(\mathbf{w})\|_2 + \epsilon'\,.$$

Therefore, for $\epsilon' \;.\; \lambda(\sqrt{\epsilon} + \|\mathbf{v} - \mathbf{w}\|_2)$ we obtain that $\|\mathbf{g}(\mathbf{w})\|_2 = O(\lambda(\sqrt{\epsilon} + \|\mathbf{v} - \mathbf{w}\|_2))$. We next show that the empirical gradient also points to the direction of $\mathbf{w} - \mathbf{w}^*$. It holds that

$$\mathbf{g}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \geq \nabla F(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) - \epsilon'\|\mathbf{w}-\mathbf{w}^*\|_2 \geq ((1/2)\alpha^2 LR^4\|\mathbf{w}-\mathbf{w}^*\|_2 - \epsilon')\|\mathbf{w}-\mathbf{w}^*\|_2\,.$$

Therefore, we need to choose $\epsilon' = O(\alpha^2 LR^4\|\mathbf{w}-\mathbf{w}^*\|_2)$. Using that $\|\mathbf{w}-\mathbf{w}^*\|_2 \geq c'\lambda\sqrt{\epsilon}/(LR^4\alpha^2)$, we have that the estimated gradient field of F satisfies the assumptions of Lemma 54 with $B = O(\lambda)$, $\alpha_1 = c'\sqrt{\epsilon}/(LR^4)\lambda/\alpha^2$ and $\alpha_2 = (1/2)\alpha^2 LR^4$. Conditioning that all the empirical gradients used by the gradient descent satisfy the error bound with ϵ' as above from Lemma 54 we obtain that with step size $\eta = \mathrm{poly}(LR\alpha/\lambda)$ after $T = \mathrm{poly}(1/(\alpha LR))W^2\log(1/\alpha_1)$ iterations we will have that $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq \mathrm{poly}(1/(LR))\lambda/\alpha^2\sqrt{\epsilon}$ which implies that

$$\begin{aligned}
F(\mathbf{w}^{(T)}) &\leq 2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2] + 2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - \sigma(\mathbf{w}^{(T)}\cdot\mathbf{x}))^2] \\
&= \mathrm{poly}(1/(LR))\;(\lambda/\alpha)^4\;O(\epsilon),
\end{aligned}$$

where we used the fact that σ is λ-Lipschitz and that the $\mathbf{x}$-marginal of D is isotropic. Since we have to do a union bound over all T iterations, and we draw fresh samples in each round, we need to divide δ by T in each round, hence, the total sample complexity is $N = \mathrm{poly}(\lambda/(\alpha LR))\,\widetilde{O}(dW^2\log^3(1/\epsilon)\log(1/\delta))$ and the runtime $\mathrm{poly}(\lambda/(\alpha LR))\log(1/\epsilon)N$. ■

C.3. Proof of Claim 55

We restate and prove the following claim.

Claim 56 *Let* $\mathbf{w}$ *be a vector with* $\|\mathbf{w}-\mathbf{w}^*\| \geq \sqrt{\epsilon}$. *Using* $N = \widetilde{O}(d\lambda^4\|\mathbf{w}-\mathbf{w}^*\|_2^2 W^2 \max(\log^2(1/\epsilon),1)/(L\epsilon')^2)$ *samples, we can compute an empirical gradient* $\mathbf{g}(\mathbf{w})$ *such that* $\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \epsilon$ *with probability* $1-\delta$.

Proof We start by bounding from above the variance in every direction. Let $\Sigma_i = \mathbf{E}[((\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}_i)^2]$, we have that

$$\begin{aligned}\Sigma_i = \mathbf{E}[((\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}_j)^2] &\leq \lambda^2\,\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2\mathbf{x}_j^2] \\ &\leq 2\lambda^2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_j^2] + 2\lambda^2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - \sigma(\mathbf{w}\cdot\mathbf{x}))^2\mathbf{x}_j^2]\,. \\ &\leq 2\lambda^2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_j^2] + 2\lambda^4\|\mathbf{w}^* - \mathbf{w}\|_2^2 \max_{\|\mathbf{u}\|_2=1}\mathbf{E}[(\mathbf{u}\cdot\mathbf{x})^2\mathbf{x}_j^2]\,.\end{aligned}$$

To bound the term $\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_j^2]$, we have that

$$\begin{aligned}\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_j^2] &= \mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_j^2\mathbf{1}\{\mathbf{x}_j \leq M\}] + \mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_j^2\mathbf{1}\{\mathbf{x}_j \geq M\}] \\ &\leq M^2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2] + 4M^2\,\mathbf{E}[\mathbf{x}_j^2\mathbf{1}\{\mathbf{x}_j \geq M\}] \\ &\lesssim M^2\epsilon + 4M^2\,\mathbf{E}[\mathbf{x}_j^2\mathbf{1}\{\mathbf{x}_j \geq M'\}] \;\lesssim\; \epsilon M^2\,.\end{aligned}$$

Moreover, note that $\sqrt{\epsilon} \leq \|\mathbf{w}-\mathbf{w}^*\|_2$, hence $\Sigma_i \lesssim \lambda^4\|\mathbf{w}-\mathbf{w}^*\|_2^2 M^2$. From Markov's inequality, we have that for each $j \leq d$, it holds

$$\begin{aligned}&\mathbf{Pr}\left[\left|\sum_{i=1}^N \frac{1}{N}(\sigma(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})\sigma'(\mathbf{w}\cdot\mathbf{x}^{(i)})\mathbf{x}_j^{(i)} - \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}_j]\right| \geq \epsilon'/\sqrt{d}\right] \\ &\leq \frac{d}{N\epsilon'^2}\,\mathbf{E}[((\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}_j)^2] \leq \frac{d\|\mathbf{w}-\mathbf{w}^*\|_2^2\lambda^4 M^2}{N\epsilon'^2}\,.\end{aligned}$$

Hence, with $N = O(\frac{d\lambda^4\|\mathbf{w}-\mathbf{w}^*\|_2^2 M^2}{\epsilon'^2})$ samples, we can get an $\epsilon'/\sqrt{d}$-approximation to the i-th coordinate of the gradient $(\nabla F(\mathbf{w}))_i$ with constant probability, and by using a standard boosting procedure we can boost the probability to $1-\delta$ with a multiplicative overhead of $O(\log(1/\delta))$ samples. Finally, doing a union bound over all coordinates $j \in \{1,\ldots,d\}$ we obtain that $N = \widetilde{O}(d\lambda^4\|\mathbf{w}-\mathbf{w}^*\|_2^2 W^2 \max(\log^2(1/\epsilon),1)/(L^2\epsilon'^2))$ samples suffice. ∎

C.4. Proof of Lemma 50

We restate and prove the following lemma.

Lemma 57 (Gradient Field Distance Reduction) *Let* $Z_1 > Z_0 \geq 1$. *Let* $\mathbf{g}: \mathbb{R}^d \mapsto \mathbb{R}^d$ *be a vector field with* $\|\mathbf{g}(\mathbf{w})\|_2 \leq B$ *for every* $\mathbf{w} \in \mathbb{R}^d$ *with* $\|\mathbf{w}\|_2 \leq 2Z_1$. *We assume that* $\mathbf{g}$ *satisfies the following properties with respect to some unknown target vector* $\mathbf{v}$*:*

1. *If* $\|\mathbf{w}\|_2 \leq Z_0$ *and* $\|\mathbf{w}-\mathbf{v}\|_2 \geq \alpha_1$ *then it holds that* $\mathbf{g}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{v}) \geq \alpha_2\|\mathbf{w}-\mathbf{v}\|_2$, *for* $\alpha_1, \alpha_2 > 0$.

2. *If* $Z_0 < \|\mathbf{w}\|_2 \leq Z_1$ *and* $\|\mathbf{w}-\mathbf{v}\|_{\mathbf{w}} \geq \beta_1$ *it holds that* $\mathbf{g}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{v}) \geq \beta_2\|\mathbf{w}-\mathbf{v}\|_{\mathbf{w}}$*, for* $\beta_1, \beta_2 > 0$.

3. *For some* $\zeta \in (0,1), \gamma > 0$*, we have that if* $\|\mathbf{w}\|_2 \geq \zeta Z_1$*, it holds that* $\mathbf{g}(\mathbf{w})\cdot\mathbf{w} \geq \gamma\|\mathbf{w}\|_{\mathbf{w}}$.

We consider the update rule $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta\mathbf{g}(\mathbf{w}^{(t)})$ *initialized with* $\mathbf{w}^{(0)} = \mathbf{0}$ *and step size*

$$\eta = \frac{1}{B^2}\min\left(\alpha_1\alpha_2, \frac{\beta_1\beta_2}{Z_1^{3/2}}, \frac{2\gamma}{Z_1}, (1-\zeta)Z_1 B\right)$$

Let T *be any integer larger than* $\left\lceil \frac{\mathsf{kvk}^2}{\eta\min(\alpha_1\alpha_2, \beta_1\beta_2/Z_1^{3/2})} \right\rceil$*. We have that* $\|\mathbf{w}^{(T)}\|_2 \leq Z_1$*. Moreover, if* $\|\mathbf{w}^{(T)}\|_2 \leq Z_0$ *it holds that* $\|\mathbf{w}^{(T)}-\mathbf{v}\|_2 \leq \eta B + \max(\alpha_1, (2Z_0)^{3/2}\beta_1)$ *and if* $\|\mathbf{w}^{(T)}\|_2 > Z_0$ *we have* $\|\mathbf{w}^{(T)}-\mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \sqrt{2}\eta B + \max(\sqrt{2}\alpha_1, e^{3Z_1^{3/2}\eta B}\beta_1)$.

Proof We first show by induction that for every $t \geq 0$ it holds that $\|\mathbf{w}^{(t)}\|_2 \leq Z_1$. For $t = 0$ we have $\mathbf{w}^{(0)} = \mathbf{0}$ so the claim holds. Assume first that $\|\mathbf{w}^{(t)}\|_2 \leq \zeta Z_1$. Then

$$\|\mathbf{w}^{(t+1)}\|_2 \leq \|\mathbf{w}^{(t)}\|_2 + \eta B \leq \zeta R_1 + (1-\zeta)Z_1 = Z_1,$$

where we used the fact that $\eta B \leq (1-\zeta)Z_1$. Now, if $\zeta Z_1 \leq \|\mathbf{w}^{(t)}\|_2 \leq Z_1$ by assumption 3 of the vector field $\mathbf{g}$, we have that

$$\begin{aligned}\|\mathbf{w}^{(t+1)}\|_2^2 &= \|\mathbf{w}^{(t)}\|_2^2 - 2\eta\mathbf{g}(\mathbf{w}^t)\cdot\mathbf{w}^{(t)} - \eta^2\|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 \\ &\leq \|\mathbf{w}^{(t)}\|_2^2 - 2\eta\gamma/\sqrt{Z_1} - \eta^2 B^2 .\end{aligned}$$

Since $\eta \leq 2\gamma/(B^2\sqrt{Z_1})$, we have that $-2\eta\gamma/\sqrt{Z_1} - \eta^2B^2 \leq 0$, and therefore $\|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 \leq Z_1^2$.

Now that we have that $\|\mathbf{w}^{(t)}\|_2$ is always smaller than Z_1, we know that the vector field $\mathbf{g}(\mathbf{w}^{(t)})$ has non-trivial component on the direction $\mathbf{w}-\mathbf{v}$, i.e., either condition 1 or condition 2 is true. We next show that when condition 1 or 2 of the Lemma 50 hold, we can improve the distance of $\mathbf{w}^{(t)}$ and the target $\mathbf{v}$. We first assume that $\|\mathbf{w}^{(t)}\|_2 \leq Z_0$ and that $\|\mathbf{w}^{(t)}-\mathbf{v}\|_2 \geq \alpha_1$. Then by assumption 1 of the lemma we have that it holds $\mathbf{g}(\mathbf{w}^{(t)})\cdot(\mathbf{w}^{(t)}-\mathbf{v}) \geq \alpha_2\|\mathbf{w}^{(t)}-\mathbf{v}\|_2$. We have

$$\begin{aligned}\|\mathbf{w}^{(t)}-\mathbf{v}\|_2^2 - \|\mathbf{w}^{(t+1)}-\mathbf{v}\|_2^2 &= 2\eta\mathbf{g}(\mathbf{w}^{(t)})\cdot(\mathbf{w}^{(t)}-\mathbf{v}) - \eta^2\|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 \\ &\geq 2\eta\alpha_1\alpha_2 - \eta^2\|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 \\ &\geq \eta\alpha_1\alpha_2 ,\end{aligned}$$

where for the last inequality we used the fact that $\|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 \leq B^2$ and that $\eta \leq \alpha_1\alpha_2/B^2$. On the other hand, if $\|\mathbf{w}^{(t)}\|_2 \leq Z_0$ and $\|\mathbf{w}^{(t)}-\mathbf{v}\|_2 \geq \beta_1$, then by assumption 2 of the lemma we have that it holds $\mathbf{g}(\mathbf{w}^{(t)})\cdot(\mathbf{w}^{(t)}-\mathbf{v}) \geq \beta_2\|\mathbf{w}-\mathbf{v}\|_{\mathbf{w}^{(t)}}$. Similarly to the previous case, we then have

$$\|\mathbf{w}^{(t)}-\mathbf{v}\|_2^2 - \|\mathbf{w}^{(t+1)}-\mathbf{v}\|_2^2 \geq 2\eta\beta_2\|\mathbf{w}^{(t)}-\mathbf{v}\|_{\mathbf{w}^{(t)}} - \eta^2\|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 .$$

We next bound from below the norm $\|\cdot\|_{\mathbf{w}}$ by the ℓ_2 norm. The following rough estimate suffices.

Claim 58 *For every* $\mathbf{x} \in \mathbb{R}^d$ *it holds*

$$\min\left(\frac{1}{\|\mathbf{w}\|_2^{3/2}}, \frac{1}{\|\mathbf{w}\|_2^{1/2}}\right)\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_{\mathbf{w}} \leq \sqrt{2}\max\left(\frac{1}{\|\mathbf{w}\|_2^{3/2}}, \frac{1}{\|\mathbf{w}\|_2^{1/2}}\right)\|\mathbf{x}\|_2 .$$

Proof The claim follows directly from the definition of the norm $\|\cdot\|_{\mathbf{w}}$ and the inequality $\sqrt{a^2+b^2} \leq a+b \leq \sqrt{2}\sqrt{a^2+b^2}$ that holds for all $a, b \geq 0$. ■

Using the above bounds and the fact that $Z_1 \geq 1$, we conclude that $\|\mathbf{w}^{(t)} - \mathbf{v}\|_2^2 - \|\mathbf{w}^{(t+1)} - \mathbf{v}\|_2^2 \geq \eta\beta_1\beta_2/Z_1^{3/2}$, where we used the assumption that $\eta \leq \beta_1\beta_2/(Z_1^{3/2}B^2)$. Recall that we have set $\eta = 1/B^2 \min(\alpha_1\alpha_2, \beta_1\beta_2/Z_1^{3/2}, 2\gamma/Z_1, (1-\zeta)Z_1 B)$ which implies that at every iteration where either $\|\mathbf{w}^{(t)}\|_2 \leq Z_0$ and $\|\mathbf{w}^{(t)} - \mathbf{v}\|_2 \geq \alpha_1$, or $\|\mathbf{w}^{(t)}\|_2 > Z_0$ and $\|\mathbf{w}^{(t)} - \mathbf{v}\|_{\mathbf{w}^{(t)}} \geq \alpha_1$ we have that $\|\mathbf{w}^{(t)} - \mathbf{v}\|_2^2 - \|\mathbf{w}^{(t+1)} - \mathbf{v}\|_2^2 \geq \eta \min(\alpha_1\alpha_2, \beta_1\beta_2/Z_1^{3/2})$. Let us denote by T the first iteration such that the distance reduction stops happening. Since we initialize at $\mathbf{w}^{(0)} = \mathbf{0}$ the first time that either of the above condition stops being true, namely, T can be at most $\left\lceil \frac{\|\mathbf{v}\|^2}{\eta \min(\alpha_1\alpha_2, \beta_1\beta_2/Z_1^{3/2})} \right\rceil$.

After either $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq \alpha_1$ or $\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \beta_1$ we stop having the guarantee that updating the guess with $\mathbf{g}(\mathbf{w}^{(T)})$ will decrease the distance, and therefore we may move to the wrong direction. However, since our step size η is small doing a small step in the wrong direction cannot increase the distance of $\mathbf{w}^{(T+1)}$ and $\mathbf{v}$ by a lot. We next show that even if we continue updating after the iteration T, we cannot make the distance of $\mathbf{w}^{(T)}$ and the target vector $\mathbf{v}$ much larger. We first assume that at iteration T we have $\|\mathbf{w}^{(T)}\|_2 \leq Z_0$, and therefore, it must be the case that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq \alpha_1$. Notice that it suffices to show that after one iteration the claim holds, since after $\|\mathbf{w}^{(T+1)}\|_2$ grows larger than α_1 the distance of $\mathbf{w}$ will start to decrease again. For the next iteration, using the bound $\|\mathbf{g}(\mathbf{w}^{(T)})\|_2 \leq B$, we obtain that $\|\mathbf{w}^{(T+1)} - \mathbf{w}^{(T)}\|_2 \leq \eta B$. Now if $\|\mathbf{w}^{(T+1)}\|_2 \leq Z_0$, using the triangle inequality for the ℓ_2 norm, we have that $\|\mathbf{w}^{(T+1)} - \mathbf{v}\|_2 \leq \alpha_1 + \eta B$. On the other hand, if $\|\mathbf{w}^{(T+1)}\|_2 > Z_0$, using the triangle inequality of the norm $\|\cdot\|_{\mathbf{w}}$ we obtain

$$\begin{aligned} \|\mathbf{w}^{(T+1)} - \mathbf{v}\|_{\mathbf{w}^{(T+1)}} &\leq \|\mathbf{w}^{(T+1)} - \mathbf{w}^{(T)}\|_{\mathbf{w}^{(T+1)}} + \|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T+1)}} \\ &\leq \sqrt{2}(\eta B + \alpha_1)\,, \end{aligned} \tag{13}$$

where we used Claim 58, the fact that $\mathbf{w}^{(T+1)} - \mathbf{w}^{(T)} = \eta\mathbf{g}(\mathbf{w}^{(T)})$ and that $\|\mathbf{w}^{(T+1)}\|_2 > Z_0 \geq 1$. Next we consider the case $\|\mathbf{w}^{(T)}\|_2 > Z_0$ and $\|\mathbf{w}^{(T+1)}\|_2 > Z_0$. From Claim 58, we obtain that $\|\mathbf{w}^{(T+1)} - \mathbf{w}^{(T)}\|_{\mathbf{w}^{(T+1)}} \leq \sqrt{2}\eta B$. We prove the next lemma that bounds the ratio between two weighted euclidean norms with different bases.

Lemma 59 *Let $\mathbf{u}, \mathbf{v}, \mathbf{x} \in \mathbb{R}^d$ be non-zero vectors with $\|\mathbf{u}\|_2, \|\mathbf{v}\|_2 \leq Q$ for some $Q \geq 1$. Then it holds that*

$$\frac{\|\mathbf{x}\|_{\mathbf{u}}}{\|\mathbf{x}\|_{\mathbf{v}}} \leq \exp\left(Q^{3/2}\Big(4\theta(\mathbf{u},\mathbf{v})\Big(\frac{1}{\|\mathbf{v}\|^{3/2}} + \frac{1}{\|\mathbf{v}\|_2^{1/2}}\Big) + \Delta(\mathbf{u},\mathbf{v})\Big)\right)\,,$$

where $\Delta(\mathbf{u},\mathbf{v}) = \left|\frac{1}{\|\mathbf{v}\|\|\mathbf{x}\|_2^{3/2}} - \frac{1}{\|\mathbf{u}\|\|\mathbf{x}\|_2^{3/2}}\right| + \left|\frac{1}{\|\mathbf{v}\|\|\mathbf{x}\|_2^{1/2}} - \frac{1}{\|\mathbf{u}\|\|\mathbf{x}\|_2^{1/2}}\right|$.

Proof We first observe that, $\|\lambda\mathbf{x}\|_{\mathbf{u}} = |\lambda|\|\mathbf{x}\|_{\mathbf{u}}$, it suffices to consider $\mathbf{x}$ with unit norm $\|\mathbf{x}\|_2 = 1$. Moreover, from Claim 58 we have that since $\|\mathbf{v}\|_2, \|\mathbf{u}\|_2 \leq Q$ it holds that both $\|\mathbf{x}\|_{\mathbf{u}}, \|\mathbf{x}\|_{\mathbf{v}}$ are larger than $1/Q^{3/2}$. Next, using the fact that the function $t \mapsto \log(t)$ is $1/t$-Lipschitz, we obtain that $\log\left(\frac{\|\mathbf{x}\|\|\mathbf{u}\|}{\|\mathbf{x}\|\|\mathbf{v}\|}\right) \leq Q^{3/2}|\|\mathbf{x}\|_{\mathbf{u}} - \|\mathbf{x}\|_{\mathbf{v}}|$. Therefore, it suffices to bound the difference of the two norms. We have that

$$|\|\mathbf{x}\|_{\mathbf{u}} - \|\mathbf{x}\|_{\mathbf{v}}| \leq \left|\frac{\|\mathrm{proj}_{\mathbf{u}}\mathbf{x}\|_2}{\|\mathbf{u}\|^{3/2}} - \frac{\|\mathrm{proj}_{\mathbf{v}}\mathbf{x}\|_2}{\|\mathbf{v}\|^{3/2}}\right| + \left|\frac{\|\mathrm{proj}_{\mathbf{u}^?}\mathbf{x}\|_2}{\|\mathbf{u}\|^{1/2}} - \frac{\|\mathrm{proj}_{\mathbf{v}^?}\mathbf{x}\|_2}{\|\mathbf{v}\|^{1/2}}\right|\,.$$

We first bound the term $\left|\frac{k\mathrm{proj}_{\mathbf{u}}\mathbf{x}k_2}{k\mathbf{u}k^{3/2}} - \frac{k\mathrm{proj}_{\mathbf{v}}\mathbf{x}k_2}{k\mathbf{v}k^{3/2}}\right| \leq \left|\frac{k\mathrm{proj}_{\mathbf{u}}\mathbf{x}k_2}{k\mathbf{v}k^{3/2}} - \frac{k\mathrm{proj}_{\mathbf{v}}\mathbf{x}k_2}{k\mathbf{v}k^{3/2}}\right| + \left|\frac{1}{k\mathbf{u}k^{3/2}} - \frac{1}{k\mathbf{v}k^{3/2}}\right|$, since $\|\mathrm{proj}_{\mathbf{v}}\mathbf{x}\|_2 \leq \|\mathbf{x}\|_2 = 1$. Moreover, by Cauchy-Schwarz inequality, we have that $\|\mathrm{proj}_{\mathbf{v}}\mathbf{x} - \mathrm{proj}_{\mathbf{u}}\mathbf{x}\|_2 \leq 2\theta(v,u)$. Similarly, we can bound from above the term $\|\mathrm{proj}_{\mathbf{v}^?}\mathbf{x}\mathrm{proj}_{\mathbf{u}^?}\|_2$ by $2\theta(\mathbf{v},u)$. The bound follows. ■

We will now use Lemma 59. We denote $\theta(\mathbf{w}^{(T)},\mathbf{w}^{(T+1)})$ the angle between $\mathbf{w}^{(T)},\mathbf{w}^{(T+1)}$ and by $\Delta(\mathbf{w}^{(T)},\mathbf{w}^{(T+1)})$ the corresponding difference defined in Lemma 59. Since $\|\mathbf{w}^{(T)}\|_2, \|\mathbf{w}^{(T+1)}\|_2 \geq Z_0 \geq 1$, we have that $\theta(\mathbf{w}^{(T)},\mathbf{w}^{(T+1)}) \leq \|\mathbf{w}^{(T)} - \mathbf{w}^{(T+1)}\|_2 \leq \eta B$. Moreover, using the fact that the mappings $\mathbf{w} \mapsto \|\mathbf{w}\|_2^{3/2}$ and $\mathbf{w} \mapsto \|\mathbf{w}\|_2^{1/2}$ defined for $\|\mathbf{w}\|_2 \geq 1$ are $3/2$ and $1/2$-Lipschitz respectively, we obtain that $\Delta(\mathbf{w}^{(T)},\mathbf{w}^{(T+1)}) \leq \|\mathbf{w}^{(T)} - \mathbf{w}^{(T+1)}\|_2 \leq \eta B$. Using Lemma 59, we conclude that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T+1)}} \leq e^{3Z_1^{3/2}\eta B}\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T)}}$. Using the triangle inequality similarly to Equation (13) and the fact that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \beta_1$, we obtain

$$\|\mathbf{w}^{(T+1)} - \mathbf{v}\|_{\mathbf{w}^{(T+1)}} \leq \sqrt{2}\eta B + e^{3Z_1^{3/2}\eta B}\beta_1 .$$

Finally, we have to consider the case where $\|\mathbf{w}^{(T)}\|_2 > Z_0$ and $\|\mathbf{w}^{(T+1)}\|_2 \leq Z_0$. Using once more the triangle inequality and Claim 58, we obtain $\|\mathbf{w}^{(T+1)} - \mathbf{v}\|_2 \leq \eta B + \|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq \eta B + (Z_0 + \eta B)^{3/2}\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \eta B + (2Z_0)^{3/2}\beta_1$, where we used the fact that $\eta B \leq 1$ and $Z_0 \geq 1$. Combining the bounds for the two cases where $\|\mathbf{w}^{(T+1)}\|_2 \leq Z_0$, we obtain that in this case it holds $\|\mathbf{w}^{(T+1)} - \mathbf{v}\|_2 \leq \eta B + \max(\alpha_1, (2Z_0)^{3/2}\beta_1)$. Similarly, when $\|\mathbf{w}^{(T+1)}\|_2 > Z_0$ we have $\|\mathbf{w}^{(T+1)} - \mathbf{v}\|_{\mathbf{w}^{(T+1)}} \leq \sqrt{2}\eta B + \max(\sqrt{2}\alpha_1, e^{3Z_1^{3/2}\eta B}\beta_1)$.

■

C.5. Proof of Claim 51

We restate and prove the following claim.

Claim 60 (Sample Complexity of Gradient Estimation) *Fix $B, \epsilon' > 0$ with $\epsilon' \leq 1/\sqrt{B}$. Using $N = \widetilde{O}((d/\epsilon'^2)\mathrm{poly}(\xi/(L\mu))\log(B/\delta))$ samples from D, we can compute an empirical gradient field $\mathbf{g}(\mathbf{w})$ such that for all $\mathbf{w}$ with $\|\mathbf{w}\| \leq B$, it holds $\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \epsilon'$ and $\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_{\mathbf{w}} \leq \epsilon'$ with probability $1-\delta$.*

Proof The result follows from the fact that the gradient random variable $\nabla F(\mathbf{w})$ is sub-exponential and therefore, its empirical estimate achieves fast convergence rates. We draw use N samples $(\mathbf{x}^{(i)}, y^{(i)})$ from D and form the standard empirical estimate

$$\mathbf{g}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{m}(\sigma(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})\mathbf{x}^{(i)} .$$

We first prove the estimation error in the dual norm of $\|\cdot\|_{\mathbf{w}}$. Recall that the dual norm is equal to $\|\mathbf{u}\|_{\mathbf{w}} = \max(\|\mathrm{proj}_{\mathbf{w}}\mathbf{u}\|_2\|\mathbf{w}\|_2^{3/2}, \|\mathrm{proj}_{\mathbf{w}^?}\mathbf{u}\|_2\|\mathbf{w}\|_2^{1/2})$. Recall that $\sigma'(\mathbf{w}\cdot\mathbf{x}) \geq 0$ and $\sigma'(\mathbf{w}\cdot\mathbf{x}) \leq \xi e^{\mu j\mathbf{w}\ \mathbf{x}j}$ for all $\mathbf{x}$. We first show that the distribution of $\mathbf{x}(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})$ is sub-exponential. Fix any unit direction $\mathbf{v}$, we show an upper bound on the tail probability $|\mathbf{v}\cdot\mathbf{x}| \geq t$. We first consider the case $\mathbf{v}\cdot\mathbf{w} = 0$ and without loss of generality we may assume that $\mathbf{v}$ is parallel to $\mathbf{e}_1$ and $\mathbf{w}$ is

parallel to $\mathbf{e}_2$. In the following calculations we repeatedly use the properties of (L, R)-well-behaved distributions and (τ, μ, ξ)-sigmoidal activations. We have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\ D}[\mathbf{1}\{|\mathbf{v}\cdot\mathbf{x}(\sigma(\mathbf{w}\cdot\mathbf{x})-y)|\geq t\}\sigma'(\mathbf{w}\cdot\mathbf{x})] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\ D}[\mathbf{1}\{|\mathbf{v}\cdot\mathbf{x}|\geq t(\xi/\mu)\}\sigma'(\mathbf{w}\cdot\mathbf{x})] \\
&\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\ D}[\mathbf{1}\{|\mathbf{v}\cdot\mathbf{x}|\geq t(\xi/\mu)\}\xi e^{\ \mu j\mathbf{w}\ \mathbf{x}j}] \leq \frac{1}{L}\int_{1}^{1}\int_{1}^{1}\mathbf{1}\{|\mathbf{x}_1|\geq t(\mu/\xi)\}e^{\ \mu k\mathbf{w}k_2 j\mathbf{x}_2 j}e^{\ L\sqrt{\mathbf{x}_1^2+\mathbf{x}_2^2}}d\mathbf{x}_1 d\mathbf{x}_2 \\
&\leq \frac{1}{L}\int_{1}^{1}\mathbf{1}\{|\mathbf{x}_1|\geq t(\xi/\mu)\}e^{\ Lj\mathbf{x}_1 j/2}d\mathbf{x}_1\int_{1}^{1}e^{\ \mu k\mathbf{w}k_2 j\mathbf{x}_2 j}e^{\ Lj\mathbf{x}_2 j/2}d\mathbf{x}_2 = \frac{2}{L}\frac{e^{\ L/2t\mu/\xi}}{L/2}\frac{1}{\mu\|\mathbf{w}\|_2+L/2} \\
&\cdot\ \frac{1}{L^2\mu(\|\mathbf{w}\|_2+1)}e^{\ L\mu/(2\xi)\ t},
\end{aligned}$$

where we used again the fact that $|\sigma(\mathbf{w}\cdot\mathbf{x})-y|\leq\xi/\mu$, the upper bound on the derivative of sigmoidal activations, see Definition 3 and the fact that the density of any 2 dimensional projection of $D_{\mathbf{x}}$ is upper bounded by $(1/L)\exp(-L\|\mathbf{x}\|_2)$, see Definition 2. Moreover, using the same properties as above we have that the variance of $\nabla F(\mathbf{w}) = \mathbf{x}(y-\sigma(\mathbf{w}\cdot\mathbf{x})\sigma'(\mathbf{w}\cdot\mathbf{x})$ along the direction $\mathbf{v}$ is at most

$$\begin{aligned}
\mathbf{Var}[\nabla F(\mathbf{w})\cdot\mathbf{v}] &\leq \left(\frac{\xi}{\mu}\right)^2 \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{v}\cdot\mathbf{x})^2] \\
&\leq \left(\frac{\xi}{\mu}\right)^2\frac{1}{L}\int_{1}^{+1}\int_{1}^{+1}e^{\ L\sqrt{\mathbf{x}_1^2+\mathbf{x}_2^2}}\,\xi e^{\ 2\mu k\mathbf{w}k_2 j\mathbf{x}_1 j}\,d\mathbf{x}_1 d\mathbf{x}_2 \\
&\leq \frac{\xi^3}{\mu^2 L}\int_{1}^{+1}\mathbf{x}_2^2 e^{\ Lj\mathbf{x}_2 j}\,d\mathbf{x}_2\int_{1}^{+1}e^{\ 2\mu k\mathbf{w}k_2 j\mathbf{x}_1 j}\,d\mathbf{x}_1 \\
&\cdot\ \frac{\xi^3}{\mu^3 L^4}\frac{1}{\|\mathbf{w}\|_2}.
\end{aligned}$$

Therefore, along any direction $\mathbf{v}$ orthogonal to $\mathbf{w}$ we have that $\nabla F(\mathbf{w})$ is (σ_1^2, b_1)-sub-exponential with variance proxy $\sigma_1^2 = \mathrm{poly}(\xi/(L\mu))1/\|\mathbf{w}\|_2$ and rate $b_1 = O(\xi/(L\mu))$. Therefore, Bernstein's inequality (see, e.g., Vershynin (2018)) implies that the empirical estimate $g(\mathbf{w})$ with N samples satisfies

$$\mathbf{Pr}\left[|g(\mathbf{w})-\nabla F(\mathbf{w}))\cdot\mathbf{v}|\geq t\right]\leq 2e^{\ \frac{Nt^2/2}{\sigma_1^2+b_1 t}}.$$

We next analyze the gradient $\nabla F(\mathbf{w})$ along the direction of $\mathbf{w}$. We denote by $\widehat{\mathbf{w}}$ the unit vector that is parallel to $\mathbf{w}$. $\|\mathrm{proj}_{\mathbf{w}}g(\mathbf{w})-\mathrm{proj}_{\mathbf{w}}\nabla F(\mathbf{w})\|_2 = \|\widehat{\mathbf{w}}\cdot\mathbf{g}(\mathbf{w})\widehat{\mathbf{w}}-\widehat{\mathbf{w}}\cdot\mathbf{g}(\mathbf{w})\widehat{\mathbf{w}}\|_2 = |\widehat{\mathbf{w}}\cdot\mathbf{g}(\mathbf{w})-\widehat{\mathbf{w}}\cdot\mathbf{g}(\mathbf{w})|$. We first show the sub-exponential tail bound:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\ D}[\mathbf{1}\{|\mathbf{w}\cdot\mathbf{x}(\sigma(\mathbf{w}\cdot\mathbf{x})-y)|\geq t\}\sigma'(\mathbf{w}\cdot\mathbf{x})] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\ D}[\mathbf{1}\{|\mathbf{w}\cdot\mathbf{x}|\geq t(\xi/\mu)\}\sigma'(\mathbf{w}\cdot\mathbf{x})] \\
&\leq \frac{1}{L}\int_{1}^{1}\mathbf{1}\{|\mathbf{x}_2|\geq t(\mu/\xi)\}e^{\ \mu k\mathbf{w}k_2 j\mathbf{x}_2 j}e^{\ Lj\mathbf{x}_2 j}d\mathbf{x}_2 \\
&= \frac{1}{L(\mu\|\mathbf{w}\|_2+L)}e^{\ (\mu k\mathbf{w}k_2+L)\mu/\xi t}.
\end{aligned}$$

Similarly, we can compute the variance along the direction of $\mathbf{w}$.

$$\begin{aligned}\mathbf{Var}[\nabla F(\mathbf{w})\cdot\widehat{\mathbf{w}}] &\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\ D}[((\sigma(\mathbf{w}\cdot\mathbf{x})-y)\sigma^{\emptyset}(\mathbf{w}\cdot\mathbf{x})\widehat{\mathbf{w}}\cdot\mathbf{x})^2]\\ &\leq \left(\frac{\xi}{\mu}\right)^2 \mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[(\sigma^{\emptyset}(\mathbf{w}\cdot\mathbf{x})\widehat{\mathbf{w}}\cdot\mathbf{x})^2]\\ &\leq \frac{\xi^3}{\mu^2 L}\int_{1}^{+1} e^{\ 2\mu k\mathbf{w}k_2 jtj}t^2 dt\\ &= \frac{\xi^3}{2\mu^5 L}\frac{1}{\|\mathbf{w}\|_2^3},\end{aligned}$$

where we used the fact that $|\sigma(\mathbf{w}\cdot\mathbf{x})-y|\leq\xi/\mu$, the upper bound on the derivative of sigmoidal activations, see Definition 3 and the anti-concentration property of (L,R)-well-behaved distributions, see Definition 2.

Therefore, along the direction $\mathbf{w}$ we have that $\nabla F(\mathbf{w})$ is (σ_2^2,b_2)-sub-exponential with variance proxy $\sigma_2^2=\mathrm{poly}(\xi/(L\mu))1/\|\mathbf{w}\|_2^3$ and rate $b_1=O(\xi/(L\mu^2\|\mathbf{w}\|_2))$. Therefore, Bernstein's inequality (see, e.g., Vershynin (2018)) implies that the empirical estimate $g(\mathbf{w})$ with N samples, satisfies

$$\mathbf{Pr}\left[|(g(\mathbf{w})-\nabla F(\mathbf{w}))\cdot\widehat{\mathbf{w}}|\geq t\right]\leq 2e^{\ \frac{Nt^2/2}{\sigma_2^2+b_2 t}}.$$

Thus, we can now perform a union bound on every direction $\mathbf{u}$ and every $\|\mathbf{w}\|\leq B$ by using a $r=\mathrm{poly}(L\mu)/(\xi B)\epsilon^{\emptyset})$-net of the unit sphere and the ball of radius B (which will have size $r^{O}(d)$, see e.g., Vershynin (2018)), we obtain that along any direction $\mathbf{v}$ orthogonal to $\mathbf{w}$ it holds that $|\mathbf{v}\cdot(\mathbf{g}(\mathbf{w})-\nabla F(\mathbf{w}))|\leq\epsilon/\|\mathbf{w}\|_2^{1/2}|$ with probability at least $1-2r^{O(d)}e^{\ \frac{N\epsilon^2/2}{k\mathbf{w}k_2\sigma_1^2+k\mathbf{w}k_2^{1/2}b_1\epsilon}}$. Substituting the variance and rate values σ_1^1,b_1, we observe that $\|\mathbf{w}\|_2\sigma_1^2+\|\mathbf{w}\|_2^{1/2}b_1 t\leq\mathrm{poly}(\xi/(L\mu))$, where we used our assumption that $\epsilon^{\emptyset}\leq 1/\sqrt{B}$ (and that $\|\mathbf{w}\|_2\leq B$). Picking $N=\widetilde{O}(d/\epsilon^{\emptyset 2}\mathrm{poly}(\xi/(L\mu))\log(B/\delta))$ we obtain that the above bound holds with probability at least $1-\delta/2$. Similarly, performing the same union bound for the direction of $\mathbf{w}$ and using the sub-exponential bound with σ_2^2,b_2 and the fact that $\epsilon^{\emptyset}\leq 1/\sqrt{B}$ we again obtain that $|\mathbf{v}\cdot(\mathbf{g}(\mathbf{w})-\nabla F(\mathbf{w}))|\leq\epsilon^{\emptyset}/\|\mathbf{w}\|_2^{3/2}$ with probability at least $1-\delta/2$ with $N=\widetilde{O}(d/\epsilon^{\emptyset 2}\mathrm{poly}(\xi/(L\mu))\log(B/\delta))$ samples. The bound for the ℓ_2 distance of $\mathbf{g}(\mathbf{w})$ and $\nabla F(\mathbf{w})$ follows similarly from the above concentration bounds. ■

C.6. Proof of Lemma 54

We restate and prove the following:

Lemma 61 (Gradient Field Distance Reduction) *Let $\mathbf{g}:\mathbb{R}^d\mapsto\mathbb{R}^d$ be a vector field and fix $W,B\geq 1$. We assume that $\mathbf{g}$ satisfies the following properties with respect to some unknown target vector $\mathbf{v}$ with $\|\mathbf{v}\|_2\leq W$. Fix parameters $\alpha_1,\alpha_2>0$. For every $\mathbf{w}\in\mathbb{R}^d$ with $\|\mathbf{w}\|_2\leq 2W$ it holds that $\|g(\mathbf{w})\|_2\leq B\max(\|\mathbf{w}-\mathbf{v}\|_2,\alpha_1)$. Moreover, if $\|\mathbf{w}-\mathbf{v}\|_2\geq\alpha_1$ and $\theta(\mathbf{w},\mathbf{v})\in(0,\pi/2)$ then it holds that $\mathbf{g}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{v})\geq\alpha_2\|\mathbf{w}-\mathbf{v}\|_2^2$. We consider the update rule $\mathbf{w}^{(t+1)}\leftarrow\mathbf{w}^{(t)}-\eta\mathbf{g}(\mathbf{w}^{(t)})$ initialized with $\mathbf{w}^{(0)}=\mathbf{0}$ and step size $\eta\leq\alpha_2/B$. Let T be any integer larger than $\lceil(W^2+\log(1/\alpha_1))/(\eta\alpha_2)\rceil$, it holds that $\|\mathbf{w}^{(T)}-\mathbf{v}\|_2\leq(1+\eta B)\alpha_1$.*

Proof We first show that at every round the distance of $\mathbf{w}^{(t)}$ and the target $\mathbf{v}$ decreases. Since we initialize at $\mathbf{w} = \mathbf{0}$ (notice that in this case it holds trivially $\|\mathbf{w}^{(0)}\|_2 \leq W$) after the first update we have

$$\begin{aligned}\|\mathbf{w}^{(1)} - \mathbf{v}\|_2^2 - \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 &= -2\eta \mathbf{g}(\mathbf{w}^{(0)}) \cdot (\mathbf{w}^{(0)} - \mathbf{v}) + \eta^2 \|\mathbf{g}(\mathbf{w}^{(0)})\|_2^2 \\ &\leq -2\eta\alpha_2 \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 + \eta^2 B^2 \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 \\ &\leq -\eta\alpha_2 \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 \,,\end{aligned}$$

where we used that since $\eta \leq \alpha_2/B$ it holds that $\eta\alpha_2 - \eta^2 B^2 \geq 0$. Moreover, we observe that after the first iteration we are going to have $\mathbf{w}^{(1)} \cdot \mathbf{v} \geq 0$ (since otherwise the distance to $\mathbf{v}$ would increase). This implies that $\theta(\mathbf{w}^{(1)}, \mathbf{v}) \in (0, \pi/2)$. Observe now that it holds $\|\mathbf{w}^{(1)}\|_2 \leq 2W$ since $\mathbf{w}^{(1)}$ is closer to $\mathbf{v}$ than $\mathbf{w}^{(0)}$. Using induction, similarly to the previous case we can show that for all iterations t where $g(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{v}) \geq \alpha_2 \|\mathbf{w}^{(t)} - \mathbf{v}\|_2^2$ it holds $\|\mathbf{w}^{(t+1)}\|_2 \leq 2W$ and $\theta(\mathbf{w}^{(t+1)}, \mathbf{v}) \in (0, \pi/2)$ and that the distance to the target vector $\mathbf{v}$ decreases at every iteration:

$$\|\mathbf{w}^{(t+1)} - \mathbf{v}\|_2^2 \leq \|\mathbf{w}^{(t)} - \mathbf{v}\|_2^2 (1 - \eta\alpha_2) \leq \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 (1 - \eta\alpha_2)^t \,.$$

Let T^0 be the first iteration where the assumption $g(\mathbf{w}^{(T^0)}) \cdot (\mathbf{w}^{(T^0)} - \mathbf{v}) \geq \alpha_2 \|\mathbf{w}^{(T^0)} - \mathbf{v}\|_2^2$ does not hold. Since at each round we decrease the distance of $\mathbf{w}^{(t)}$ and $\mathbf{v}$, and we initialize at $\mathbf{0}$ we have that the number of iterations $T^0 \leq \lceil (\|\mathbf{v}\|_2^2 + \log(1/\alpha))/(\eta\alpha_2) \rceil$. Moreover, even if we continue to use the update rule after the iteration T^0 from the triangle inequality we have that

$$\|\mathbf{w}^{(T^0+1)} - \mathbf{v}\|_2 \leq \|\mathbf{w}^{(T^0+1)} - \mathbf{w}^{(T^0)}\|_2 + \|\mathbf{w}^{(T^0)} - \mathbf{v}\|_2 \leq \eta \|\mathbf{g}(\mathbf{w}^{(T^0)})\|_2 + \alpha_1 \leq (1 + \eta B)\alpha_1 \,.$$

Therefore, by induction we get that for all $T \geq T^0$ it holds that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq (1 + \eta B)\alpha_1$. ■

C.7. Proof of Claim 53

We restate and prove the following claim.

Claim 62 *Let $M = \Theta(W \max(\log(W/\epsilon), 1))$. Denote by $\mathrm{tr}(y) = \mathrm{sign}(y) \min(|y|, M)$. Then, it holds that:*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(\mathrm{tr}(y) - \sigma(\mathbf{w} \ \cdot \mathbf{x}))^2] \leq O(\epsilon) \,.$$

Proof Denote $\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{w} \ \cdot \mathbf{x}| \leq M\}$, i.e., the points $\mathbf{x}$ such that $|\mathbf{w} \ \cdot \mathbf{x}|$ is at most M. We have that for $\mathbf{x} \in \mathcal{B}$, it holds that $|\tilde{y} - \sigma(\mathbf{w} \ \cdot \mathbf{x})| \leq |y - \sigma(\mathbf{w} \ \cdot \mathbf{x})|$, where $\tilde{y}$ is equal to y truncated in $-M \leq y \leq M$. Therefore, we have that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(\tilde{y} - \sigma(\mathbf{w} \ \cdot \mathbf{x}))^2 \mathbf{1}\{\mathbf{x} \in \mathcal{B}\}] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(y - \sigma(\mathbf{w} \ \cdot \mathbf{x}))^2 \mathbf{1}\{\mathbf{x} \in \mathcal{B}\}] \leq \epsilon \,.$$

Using the triangle inequality, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(\tilde{y} - \sigma(\mathbf{w} \ \cdot \mathbf{x}))^2] &= \mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(\tilde{y} - \sigma(\mathbf{w} \ \cdot \mathbf{x}))^2 \mathbf{1}\{\mathbf{x} \in \mathcal{B}\}] + \mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(\tilde{y} - \sigma(\mathbf{w} \ \cdot \mathbf{x}))^2 \mathbf{1}\{\mathbf{x} \notin \mathcal{B}\}] \\ &\leq \epsilon + 2M^2 \mathop{\mathbf{Pr}}_{\mathbf{x} \ D_\mathbf{x}}[\mathbf{x} \notin \mathcal{B}] + 2 \mathop{\mathbf{E}}_{\mathbf{x} \ D_\mathbf{x}}[\sigma(\mathbf{w} \ \cdot \mathbf{x})^2 \mathbf{1}\{\mathbf{x} \notin \mathcal{B}\}] \,.\end{aligned}$$

To bound $\mathbf{E}_{\mathbf{x}\ D_{\mathbf{x}}}[\sigma(\mathbf{w}\ \cdot\mathbf{x})^2\mathbf{1}\{\mathbf{x}\notin\mathcal{B}\}]$, note that $\|\mathbf{w}\ \|_2\le W$ and that $|\sigma(\mathbf{w}\ \cdot\mathbf{x})|\le\lambda|\mathbf{w}\ \cdot\mathbf{x}|$. Assuming without loss of generality that $\mathbf{w}$ is parallel to $\mathbf{e}_1$ and let V be the subspace spanned by $\mathbf{w}$ and any other vector orthogonal to $\mathbf{w}$. Using that the distribution $D_{\mathbf{x}}$ is (L,R)-well-behaved, we have that it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[\sigma(\mathbf{w}\ \cdot\mathbf{x})^2\mathbf{1}\{\mathbf{x}\notin\mathcal{B}\}]\le\lambda^2W^2\mathop{\mathbf{E}}_{\mathbf{x}\ D_{\mathbf{x}}}[\mathbf{x}_1^2\mathbf{1}\{\mathbf{x}_1\ge M/W\}]\le\epsilon/2\,,$$

which follows from $\lambda^2\,\mathbf{E}_{\mathbf{x}\ D_{\mathbf{x}}}[|\mathbf{x}_1|^2\mathbf{1}\{\mathbf{x}_1\ge M/W\}]\le\lambda^2(M^2/(L^2W^2))\exp(-LM/W)\le\epsilon/2$, its proof is similar with the proof of Claim 36. It remains to bound the $\mathbf{Pr}_{\mathbf{x}\ D_{\mathbf{x}}}[\mathbf{x}\notin\mathcal{B}]$, we have

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\ D_{\mathbf{x}}}[\mathbf{x}\notin\mathcal{B}]=\int_{\mathbf{jw}\ \ \mathbf{xj@B}}\gamma_V(\mathbf{x})\mathrm{d}\mathbf{x}\le\frac{2}{L}\int_{t\ \ M/\mathbf{kw}\ \ \mathbf{k}_2}e^{\ Lt}\mathrm{d}t=\frac{2}{L^2}e^{\ LM/\mathbf{kw}\ \ \mathbf{k}_2}\le\frac{\epsilon}{2M^2}\,.$$

Hence, $(1/2)\,\mathbf{E}_{(\mathbf{x},y)\ \ D}[(\tilde{y}-\sigma(\mathbf{w}\ \cdot\mathbf{x}))^2]\ .\ \epsilon$. Therefore, by truncated the y to values $-M\le y\le M$, we increase the total error at most $\epsilon^{\prime}=O(\epsilon)$. For the rest of the proof for simplicity we abuse the notation of the symbol y for the $\tilde{y}$ and ϵ for the $\epsilon^{\prime}$. ■

Corollary 63 *Let D be an (L,R)-well-behaved distribution on $\mathbb{R}^d\times\mathbb{R}$ and $\sigma(\cdot)$ be a (τ,μ,ξ)-sigmoidal activation. Define $F(\mathbf{w})=(1/2)\,\mathbf{E}_{(\mathbf{x},y)\ \ D}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2]$ and $\mathrm{opt}=\inf_{\mathbf{w}2\mathbb{R}^d}F(\mathbf{w})$. Set $\kappa=\mathrm{poly}(LR\tau\mu/\xi)$. There exists an algorithm that given $N=\widetilde{\Theta}(d/\epsilon\log(1/\delta))\ \mathrm{poly}(1/\kappa)$ samples, runs in time $T=\mathrm{poly}(1/(\epsilon\kappa))$ and returns a vector $\widehat{\mathbf{w}}\in\mathbb{R}^d$ that, with probability $1-\delta$, satisfies*

$$F(\widehat{\mathbf{w}})\le\mathrm{poly}(1/\kappa)\ \mathrm{opt}+\epsilon\,.$$

Proof The proof of Corollary 63 is similar to the proof of Theorem 22, the main point which the two proofs diverge is that we have to guess the value ρ of the regularizer. To do this, we construct a grid of the possible values of opt, i.e., $\mathcal{G}=\{\epsilon,2\epsilon,\ldots,\Theta(\xi/\mu)\}$. We choose the upper bound to be of size $\Theta(\xi/\mu)$ because this is the maximum value that the error can get, see e.g., Fact 28. Therefore, by running the Algorithm 1, with appropriate parameters as in Theorem 22 but by using for the regularizer, each value of $\mathcal{G}$ instead of opt. There exist a $t\in\mathcal{G}$ such that $|\mathrm{opt}-t|\le\epsilon$, therefore from Proposition 14, this will converge to a stationary point with $(1+\epsilon)\mathrm{poly}(1/k)\mathrm{opt}$, see e.g., Proposition 14 for $\Lambda=1+\epsilon$. Hence, by running Algorithm 1 for each value in $\mathcal{G}$, we output a list U that contains $O(\xi/(\mu\epsilon))$ different hypothesis. We construct the empirical D_N of D by taking N samples. We need enough samples, such that for each $\mathbf{w}\in U$, it will hold $\widehat{F}(\mathbf{w})=\mathbf{E}_{(\mathbf{x},y)\ \ D_N}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2]\le 2\,\mathbf{E}_{(\mathbf{x},y)\ \ D}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2]$. From Markov's inequality we have that

$$\mathbf{Pr}[|\widehat{F}(\mathbf{w})-F(\mathbf{w})|\ge F(\mathbf{w})/2]\le 4\frac{\mathbf{E}[\widehat{F}(\mathbf{w})^2]}{F(\mathbf{w})^2}\le\frac{\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^4]}{NF(\mathbf{w})^2}\,.$$

Note that because $y,\sigma(\mathbf{w}\cdot\mathbf{x})$ is bounded by $\Theta(\xi/\mu)$, we have that $\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^4]\le\mathrm{poly}(\xi/\mu)\,\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2]$, hence we have that

$$\mathbf{Pr}[|\widehat{F}(\mathbf{w})-F(\mathbf{w})|\ge F(\mathbf{w})/2]\ .\ \mathrm{poly}(\xi/\mu)\frac{1}{NF(\mathbf{w})}\ .\ \mathrm{poly}(\xi/\mu)\frac{1}{N\epsilon}\,,$$

therefore by choosing $N = \mathrm{poly}(\xi/\mu)(1/\epsilon)$, we have with probability at least $2/3$ that $\widehat{F}(\mathbf{w}) \leq 2F(\mathbf{w}) + \epsilon$ and by standard boosting procedure we can increase the probability to $1 - \delta''$ with $O(\log(1/\delta''))$ samples. Hence, by choosing $\delta'' = \delta/|U|$, we have with probability $1 - \delta$, that for $\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w} 2 U} \widehat{F}(\mathbf{w})$, it holds that $F(\hat{\mathbf{w}}) \ . \ \mathrm{opt} + \epsilon$. ■

Appendix D. Learning Halfspaces using the Ramp Activation

In this section we show that by optimizing as a surrogate function the ramp activation, we can find a $\hat{\mathbf{w}} \in \mathbb{R}^d$, that gets error $\mathbf{E}_{(\mathbf{x},y) \ D}[(\mathrm{sign}(\hat{\mathbf{w}} \cdot \mathbf{x}) - y)^2] = O(\epsilon)$. Specifically, we show the following corollary of Corollary 63.

Corollary 64 *Let D be an (L, R)-well-behaved distribution on $\mathbb{R}^d \times \{\pm 1\}$. Define $F(\mathbf{w}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y) \ D}[(\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) - y)^2]$ and $\mathrm{opt} = \inf_{\mathbf{w} 2 \mathbb{R}^d} F(\mathbf{w})$. Fix $\epsilon, \delta \in (0, 1)$ and set $\kappa = \mathrm{poly}(LR)$. There exists an algorithm that given $N = \widetilde{\Theta}(d/\epsilon \log(1/\delta))\,\mathrm{poly}(1/\kappa)$ samples, runs in time $T = \mathrm{poly}(1/(\epsilon\kappa))$ and returns a vector $\widehat{\mathbf{w}} \in \mathbb{R}^d$ that, with probability $1 - \delta$, satisfies*

$$F(\widehat{\mathbf{w}}) \leq \mathrm{poly}(1/\kappa)\,\mathrm{opt} + \epsilon\,.$$

Proof We are going to show that by optimizing the ramp activation instead of the $\mathrm{sign}(\cdot)$, we can find a good candidate solution. First note that $r(t)$ is an $(1, e, 1)$-sigmoidal activation. The proof relies on the following fact:

Fact 65 *Let $\mathbf{w} \in \mathbb{R}^d$ be a unit vector. Then, for any $\mathbf{x} \in \mathbb{R}^d$, it holds that*

$$\lim_{z \ ! \ 1} r(z\mathbf{w} \cdot \mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x})\ .$$

Therefore, for any unit vector $\mathbf{w} \in \mathbb{R}^d$, from Lemma 13 there exists a $\mathbf{w}' \in \mathbb{R}^d$ with $\|\mathbf{w}'\|_2 = O(1/\epsilon)$ such that $\mathbf{E}_{\mathbf{x} \ D_{\mathbf{x}}}[(r(\mathbf{w}' \cdot \mathbf{x}) - \mathrm{sign}(\mathbf{w} \cdot \mathbf{x}))^2] \leq \epsilon$. From Corollary 63, we get that there exists an algorithm that with $N = \widetilde{\Theta}(d/\epsilon \log(1/\delta))\ \mathrm{poly}(1/\kappa)$ samples, and runtime $T = \mathrm{poly}(1/(\epsilon\kappa))$ returns $\hat{\mathbf{v}} \in \mathbb{R}^d$ with $\|\hat{\mathbf{v}}\|_2 = O(1/\epsilon)$, such that with probability $1 - \delta$, it holds

$$(1/2) \mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(r(\hat{\mathbf{v}} \cdot \mathbf{x}) - y)^2] \leq \mathrm{poly}(1/(LR))\mathrm{opt} + \epsilon\ .$$

The proof follows by noting that because $r(\hat{\mathbf{v}}) \in (-1, 1)$ and $y \in \{\pm 1\}$, it holds that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(r(\hat{\mathbf{v}} \cdot \mathbf{x}) - y)^2] \geq (1/2) \mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(\mathrm{sign}(r(\hat{\mathbf{v}} \cdot \mathbf{x})) - y)^2] = (1/2) \mathop{\mathbf{E}}_{(\mathbf{x},y) \ D}[(\mathrm{sign}(\hat{\mathbf{v}} \cdot \mathbf{x}) - y)^2]\ .$$

Hence, $F(\hat{\mathbf{v}}) \leq \mathrm{poly}(1/(LR))\mathrm{opt} + 2\epsilon$. ■

Next, we prove the following sample complexity lower bound for optimizing sigmoidal activations.

Lemma 66 (Sample Complexity Lower Bound for Sigmoidal Activations) *Fix any absolute constant $C \geq 1$ and let $\sigma(t) = 1/(1 + e^{\ t})$ be the logistic activation. Any C-approximate algorithm that learns (with success probability at least $2/3$) the logistic activation under any ϵ-corrupted D with standard normal $\mathbf{x}$-marginal requires $\Omega(d/\epsilon)$ samples from D.*

Proof Assume that $y = \mathtt{1}\{\mathbf{w} \cdot \mathbf{x} \geq 0\}$ for some unit vector $\mathbf{w}$, i.e., y is a noiseless halfspace. It is easy to see that the corresponding distribution D with Gaussian $\mathbf{x}$-marginal and y given by $\mathtt{1}\{\mathbf{w} \cdot \mathbf{x} \geq 0\}$ is ϵ-corrupted for any $\epsilon > 0$. Since $\lim_{z \to -\infty} \sigma(z\mathbf{w} \cdot \mathbf{x}) = \mathtt{1}\{\mathbf{w} \cdot \mathbf{x} \geq 0\}$, by the dominated convergence theorem we have that $\lim_{z \to -\infty} \mathbf{E}_{(\mathbf{x},y) \sim D} F^{D,\sigma}(z\mathbf{w}) = 0$. Therefore, $\inf_{\mathbf{w} \in \mathbb{R}^d} F^{D,\sigma}(\mathbf{w}) = 0$. Let $\mathcal{A}$ be any algorithm that given a sigmoidal activation σ and $N > 0$ samples from D, returns a vector $\mathbf{w}$ such that $F^{D,\sigma}(\mathbf{w}) \leq \epsilon$. We will show that $\mathbf{w}$ corresponds to a classifier with at most $2C\epsilon$ error. By noting that $\sigma(\mathbf{w}) \in (0,1)$ and $y \in \{0,1\}$, it holds that

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y) \sim D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2] &\geq (1/2) \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim D}[(\mathtt{1}\{\sigma(\mathbf{w} \cdot \mathbf{x}) \geq 1/2\}) - y)^2] \\ &= (1/2) \mathop{\mathbf{Pr}}_{(\mathbf{x},y) \sim D}[\mathtt{1}\{\sigma(\mathbf{w} \cdot \mathbf{x}) \geq 1/2\}) \neq y] \,.\end{aligned}$$

Thus, we have constructed a binary classifier that achieves disagreement at most $2C\epsilon$ with y. It is well-known that any algorithm that finds a vector $\mathbf{w} \in \mathbb{R}^d$ such that $\mathbf{Pr}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y] \leq O(\epsilon)$, needs $\Omega(d/\epsilon)$ samples (see, e.g., Long (1995)). Hence, $\mathcal{A}$ needs at least $N = \Omega(d/\epsilon)$ samples. ■