Robustly Learning Single-Index Models via Alignment Sharpness

Nikos Zarifis [*1] **Puqian Wang** [*1] **Ilias Diakonikolas** [1] **Jelena Diakonikolas** [1]

Abstract

We study the problem of learning Single-Index Models under the L_2^2 loss in the agnostic model. We give an efficient learning algorithm, achieving a constant factor approximation to the optimal loss, that succeeds under a range of distributions (including log-concave distributions) and a broad class of monotone and Lipschitz link functions. This is the first efficient constant factor approximate agnostic learner, even for Gaussian data and for any nontrivial class of link functions. Prior work for the case of unknown link function either works in the realizable setting or does not attain constant factor approximation. The main technical ingredient enabling our algorithm and analysis is a novel notion of a local error bound in optimization that we term *alignment sharpness* and that may be of broader interest.

1. Introduction

Single-index models (SIMs) (Ichimura, 1993; Hristache et al., 2001; Härdle et al., 2004; Dalalyan et al., 2008; Kalai & Sastry, 2009; Kakade et al., 2011; Dudeja & Hsu, 2018) are a classical supervised learning model extensively studied in statistics and machine learning. SIMs capture the common assumption that the target function f depends on an unknown direction $\mathbf{w}$, i.e., $f(\mathbf{x}) = u(\mathbf{w} \cdot \mathbf{x})$ for some link (a.k.a. activation) function $u : \mathbb{R} \mapsto \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$. In most settings, the link function is unknown and is assumed to satisfy certain regularity properties. Classical works (Kalai & Sastry, 2009; Kakade et al., 2011) studied the efficient learnability of SIMs for monotone and Lipschitz link functions and data distributed on the unit ball. These early algorithmic results succeed in the realizable setting (i.e., with clean labels) or in the presence of zero-mean label noise.

The focus of this work is on learning SIMs in the challenging agnostic (or adversarial label noise) model (Haussler, 1992; Kearns et al., 1994), where no assumptions are made on the labels of the examples and the goal is to compute a hypothesis that is competitive with the *best-fit* function in the class. Importantly, as will be formalized below, we will not assume a priori knowledge of the link function. In more detail, let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ and $\mathcal{L}_2(h) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(h(\mathbf{x}) - y)^2]$ be the squared loss of the hypothesis $h : \mathbb{R}^d \to \mathbb{R}$ with respect to $\mathcal{D}$. Given i.i.d. samples from $\mathcal{D}$, the goal of the learner is to output a hypothesis h with squared error competitive with OPT, where $\mathrm{OPT} = \inf_{f\in\mathcal{C}} \mathcal{L}_2(f)$ is the best attainable error by any function in the target class $\mathcal{C}$.

In the context of this paper, the class $\mathcal{C}$ above is the class of SIMs, i.e., all functions of the form $f(\mathbf{x}) = u(\mathbf{w} \cdot \mathbf{x})$ where both the weight vector $\mathbf{w}$ and the link function u are unknown. For this task to be even information-theoretically solvable, one requires some assumptions on the vector $\mathbf{w}$ and the link function u. We will assume, as is standard, that the ℓ_2-norm of $\mathbf{w}$ is bounded by a parameter W. We will similarly assume that the link function lies in a family of well-behaved functions that are monotone and satisfy certain Lipschitz properties (see Definition 1.3).

For a weight vector $\mathbf{w}$ and link function u , the L_2^2 loss of the SIM hypothesis $u(\mathbf{w} \cdot \mathbf{x})$ (defined by u and $\mathbf{w}$) is $\mathcal{L}_2(\mathbf{w}; u) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. Our problem of robustly learning SIMs is defined as follows.

Problem 1.1 (Robustly Learning Single-Index Models)**.** Fix a class of distributions $\mathcal{G}$ on $\mathbb{R}^d$ and class of link functions[1] $\mathcal{F}$. Let $\mathcal{D}$ be a distribution of labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that its $\mathbf{x}$-marginal $\mathcal{D}_\mathbf{x}$ belongs to $\mathcal{G}$. We say that an algorithm is a C-approximate proper SIM learner, for some $C \geq 1$, if given $\epsilon > 0$, $W > 0$, and i.i.d. samples from $\mathcal{D}$, the algorithm outputs a link function $\hat{u} \in \mathcal{F}$ and a vector $\widehat{\mathbf{w}} \in \mathbb{R}^d$ such that with high probability it holds $\mathcal{L}_2(\widehat{\mathbf{w}}; \hat{u}) \leq C\,\mathrm{OPT} + \epsilon$, where $\mathrm{OPT} \triangleq \min_{\|\mathbf{w}\|_2\leq W, u\in\mathcal{F}} \mathcal{L}_2(\mathbf{w}; u)$.

Throughout, we use $u^*(\mathbf{w}^* \cdot \mathbf{x})$ to denote an(y) fixed optimal solution to the learning problem, i.e., $\mathcal{L}_2(\mathbf{w}^*; u^*) = \mathrm{OPT}$.

Some comments are in order. First, Problem 1.1 does not make realizability assumptions on the distribution $\mathcal{D}$. That

[*]Equal contribution [1]Department of Computer Science, University of Wisconsin - Madison, Madison, WI, USA. Correspondence to: Puqian Wang <pwang333@wisc.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Throughout this paper, we will use the terms "link function" and "activation" interchangeably.

is, the labels are allowed to be arbitrary and the goal is to be competitive against the best-fit function in the class $\mathcal{C} = \{u(\mathbf{w} \cdot \mathbf{x}) \mid \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq W, u \in \mathcal{F}\}$. Second, our focus is on obtaining efficient learners that achieve a *constant factor approximation* to the optimum loss, i.e., where C in Problem 1.1 is a *universal constant* — independent of the dimension d and the radius W of the weight space.

Ideally, one would like an efficient learner that succeeds for all marginal distributions and achieves optimal error of $\mathrm{OPT} + \epsilon$ (corresponding to $C = 1$). Unfortunately, known computational hardness results rule out this possibility. Even for the very special case that the marginal distribution is Gaussian and the link function is known (e.g., a ReLU), there is strong evidence that any algorithm achieving error $\mathrm{OPT} + \epsilon$ requires $d^{\mathrm{poly}(1/\epsilon)}$ time (Diakonikolas et al., 2020b; Goel et al., 2020; Diakonikolas et al., 2021; 2023). Moreover, even if we relax our goal to constant factor approximation (i.e., $C = O(1)$), distributional assumptions are required both for proper (Síma, 2002; Manurangsi & Reichman, 2018) and improper learning (Diakonikolas et al., 2022a). As a consequence, algorithmic research in this area has focused on constant factor approximate learners that succeed under mild distributional assumptions.

Recent works (Diakonikolas et al., 2020a; 2022c; Awasthi et al., 2023; Wang et al., 2023) gave efficient, constant factor approximate learners, under natural distributional assumptions, for the special case of Problem 1.1 where the link function is known *a priori* (see also Frei et al. (2020)). For the general setting, the only prior algorithmic result was recently obtained in Gollakota et al. (2023). Specifically, Gollakota et al. (2023) gave an efficient algorithm that succeeds for the class of monotone 1-Lipschitz link functions and any marginal distribution with second moment bounded by λ. Their algorithm achieves L_2^2 error

$$O(W\sqrt{\lambda}\sqrt{\mathrm{OPT}}) + \epsilon \tag{1}$$

under the assumption that the labels are bounded in $[0, 1]$. The error guarantee (1) is substantially weaker — both qualitatively and quantitatively — from the goal of this paper. Firstly, the dependence on OPT scales with its square root, as opposed to linearly. Secondly, and arguably importantly, the multiplicative factor inside the big-O scales (linearly) with the diameter of the space W.

Interestingly, Gollakota et al. (2023) showed — via a hardness construction from Diakonikolas et al. (2022a) — that, under their distributional assumptions, a multiplicative dependence on W (in the error guarantee) is inherent for efficient algorithms. That is, to obtain an efficient constant factor approximation, it is necessary to restrict ourselves to distributions with *additional* structural properties. This discussion raises the following question:

Can we obtain efficient constant factor *learners for Problem 1.1 under mild distributional assumptions?*

The natural goal here is to match the guarantees of known algorithmic results for the special case of known link function (Diakonikolas et al., 2022c; Wang et al., 2023).

As our main contribution, we answer this question in the affirmative. That is, we give *the first* efficient constant factor approximation learner that succeeds under natural and broad families of distributions (including log-concave distributions) and a broad class of link functions. We emphasize that this is the first polynomial-time constant factor approximate learner even for Gaussian marginals and for any nontrivial class of link functions. Roughly speaking, our distributional assumptions require concentration and (anti)-anti-concentration (see Definition 1.2).

1.1. Overview of Results

We start by providing the distributional assumptions and family of link functions for which our algorithm succeeds.

Distributional Assumptions Our algorithm succeeds for the following class of structured distributions.

Definition 1.2 (Well-Behaved Distributions)**.** Let $L, R > 0$. Let V be any subspace in $\mathbb{R}^d$ of dimension at most 2. A distribution $\mathcal{D}_{\mathbf{x}}$ on $\mathbb{R}^d$ is called (L, R)-well-behaved if $\mathcal{D}_{\mathbf{x}}$ is isotropic and for any projection $(\mathcal{D}_{\mathbf{x}})_V$ of $\mathcal{D}_{\mathbf{x}}$ onto subspace V, the corresponding pdf γ_V on $\mathbb{R}^2$ satisfies the following:

- For all $\mathbf{x}_V \in V$ such that $\|\mathbf{x}_V\|_\infty \leq R$, $\gamma_V(\mathbf{x}_V) \geq L$ (anti-anti-concentration).
- For all $\mathbf{x}_V \in V$, $\gamma_V(\mathbf{x}_V) \leq (1/L)(e^{-L\|\mathbf{x}_V\|_2})$ (anti-concentration and concentration).

This distribution class was introduced in Diakonikolas et al. (2020c), in the context of learning linear separators with noise and has since been used in a number of prior works, including for robustly learning SIMs with known link functions (Diakonikolas et al., 2022c). The parameters L, R in Definition 1.2 are viewed as universal constants, i.e., $L, R = O(1)$. Indeed, it is known that many natural distributions, most importantly isotropic log-concave distributions, fall in this category; see, e.g., Diakonikolas et al. (2020c).

Unbounded Activations Our algorithm succeeds for a broad class of link functions that contains many well-studied activations, including ReLU. This class, defined in Diakonikolas et al. (2022c) and used in Wang et al. (2023), requires the link functions to be monotone, Lipschitz-continuous and strictly increasing in the positive region.

Definition 1.3 (Unbounded Activations)**.** Let $u : \mathbb{R} \mapsto \mathbb{R}$. Given $a, b \in \mathbb{R}$ such that $0 < a \leq b$, we say that $u(z)$ is (a, b)-well-behaved if $u(0) = 0$ and $u(z)$ is non-decreasing, b-Lipschitz-continuous, and $u(z) - u(z') \geq a(z - z')$ for all $z \geq z' \geq 0$. We denote this function class by $\mathcal{U}_{(a,b)}$.

A simplified version of our main algorithmic result is as follows (see Theorem 4.2 for a more detailed statement):

Theorem 1.4 (Main Algorithmic Result, Informal)**.** *Given Problem 1.1, where $\mathcal{G}$ is the class of (L, R)-well behaved distributions with $L, R = O(1)$ and $\mathcal{F} = \mathcal{U}_{(a,b)}$ such that $(1/a), b = O(1)$, there is an algorithm that draws $N = \mathrm{poly}(W)\tilde{O}(d/\epsilon^2)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(N, d)$ time, and outputs a hypothesis $\hat{u}(\widehat{\mathbf{w}} \cdot \mathbf{x})$ with $\hat{u} \in \mathcal{U}_{(a,b)}, \|\widehat{\mathbf{w}}\|_2 \leq W$ such that $\mathcal{L}_2(\widehat{\mathbf{w}}; \hat{u}) = C\mathrm{OPT} + \epsilon$ with high probability, where $C > 0$ is an absolute constant.*

We reiterate that the approximation factor C in Theorem 1.4 is a universal constant, independent of the dimension and diameter of the space. That is, our main result provides the first efficient learning algorithm achieving a constant factor approximation, even for the most basic case of Gaussian data and any non-trivial class of link functions.

1.2. Technical Overview

When it comes to learning SIMs in the agnostic model with target error $C\mathrm{OPT} + \epsilon$, to the best of our knowledge, all prior work that achieves such a guarantee with C being an absolute constant only applies to the special case of known link function u^*. Such results are established by proving growth conditions (local error bounds) that relate either the L_2^2 loss or a surrogate loss to (squared) distance to the set of target solutions, using assumptions about the link function and the data distribution, such as concentration and (anti-)anti-concentration (Diakonikolas et al., 2020a; 2022b; Wang et al., 2023). Among these, most relevant to our work is Wang et al. (2023), which proved a "sharpness" property for a convex surrogate function defined by

$$\mathcal{L}_{\mathrm{sur}}(\mathbf{w}; u) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\int_0^{\mathbf{w}\cdot\mathbf{x}} (u(r) - y)\,\mathrm{d}r\right], \tag{2}$$

based on assumptions about the link function that are the same as ours and distributional assumptions that are somewhat weaker but comparable to ours. Their sharpness result corresponds to guaranteeing that for vectors $\mathbf{w}$ that are not already $O(\mathrm{OPT})+\epsilon$ accurate solutions, the following holds:

$$\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}; u^*) \cdot (\mathbf{w} - \mathbf{w}^*) \gtrsim \|\mathbf{w} - \mathbf{w}^*\|_2^2, \tag{3}$$

where $\mathbf{w}^*$ is a vector that achieves error $O(\mathrm{OPT}) + \epsilon$.

One may hope that the sharpness result of Wang et al. (2023) can be generalized to the case of unknown link function and leveraged to obtain constant factor robust learners. However, as we discuss below, such direct generalizations are not possible and there are several technical challenges that had to be overcome in our work. To illustrate some of the intricacies, consider first the following example.

Example 1.5. Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{w} = (1/2)\mathbf{w}^*$, where $\mathbf{w}^*$ is an arbitrary but fixed target unit vector. Let $b > 2a$. Suppose that the link function at hand is $u(z) = bz$ and the target link function is $u^*(z) = az$. Observe that both $u, u^* \in \mathcal{U}_{(a,b)}$, as required by our model. Furthermore, suppose there is no label noise, in which case $\mathrm{OPT} = 0$. Note that the L_2^2 error of $u(\mathbf{w} \cdot \mathbf{x})$ in this case is

$$\begin{aligned}\mathcal{L}_2(\mathbf{w}; u) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(u(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \\ &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}[(b/2 - a)^2 z^2] = (b/2 - a)^2 = \Theta(1).\end{aligned}$$

However, the gradient of the surrogate loss, $\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}; u) = \mathbf{E}[(u(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))\mathbf{x}]$, is negatively correlated with $\mathbf{w} - \mathbf{w}^*$, i.e., $\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}; u) \cdot (\mathbf{w} - \mathbf{w}^*) < 0$, contrary to what we would hope for if a sharpness property as in Wang et al. (2023) were to hold. Thus, although $\mathbf{w}$ and u are both still far away from the target parameters $\mathbf{w}^*$ and u^*, the gradient of the surrogate loss cannot provide useful information about the direction in which to update $\mathbf{w}$.

What Example 1.5 demonstrates is that we cannot hope for the surrogate loss to satisfy a local error bound for an arbitrary parameter pair $(u, \mathbf{w})$ that would guide the convergence of an algorithm toward a target parameter pair $(u^*, \mathbf{w}^*)$. This seemingly insurmountable obstacle is surpassed by observing that we do not, in fact, need the surrogate loss to contain a "signal" that would guide us toward target parameters for an *arbitrary* pair $(u, \mathbf{w})$. Instead, we can restrict our attention to pairs $(u, \mathbf{w})$ satisfying that u is a "reasonably good" link function for the vector $\mathbf{w}$. Ideally, we would like to only consider link functions u that minimize the L_2^2 loss — considering that u^* must minimize the L_2^2 loss for a given, fixed $\mathbf{w}^*$ — but it is unclear how to achieve that in a statistically and computationally efficient manner. As a natural approach, we consider link functions that are the best fitting functions in an empirical distribution sense. In particular, given a sample set $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ and a parameter $\mathbf{w}$, we select a function $\hat{u}_{\mathbf{w}}$ that solves the following (convex) optimization problem:

$$\hat{u}_{\mathbf{w}} \in \mathop{\mathrm{argmin}}_{u\in\mathcal{U}_{(a,b)}} \frac{1}{m}\sum_{i=1}^m (u(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})^2. \tag{P}$$

For notational simplicity, we drop the parameter $\mathbf{w}^t$ from $\hat{u}_{\mathbf{w}^t}$ and use $\hat{u}^t$ instead. It is worth pointing out here that in general the problem of finding the best function that minimizes the L_2^2 error fails under the category of non-parametric regression, which unfortunately requires exponentially many samples (namely, $\Omega(1/\epsilon^d)$). Fortunately, in our setting, we are looking for the best function that lies in a *one-dimensional space*. Therefore, instead of looking at all possible directions, we can project all the points of the sample set S to the direction $\mathbf{w}$ and find the best fitting link function efficiently. We provide the full details for efficiently solving (P) in Appendix E.

Having set on the "best-fit" link functions in the sense of the problem (P), the next obstacle one encounters when trying to prove a "sharpness-like" result is that *neither the L_2^2 loss nor its surrogate convey information about the scale* of $\mathbf{w}$ and $\mathbf{w}^*$. This is because models determined by $u, \mathbf{w}$ and $u/c, c\mathbf{w}$ for some parameter $c > 0$ have the same value of both loss functions. Thus, it seems unlikely that a more traditional local error bound as in (3) can be established in general, for either the surrogate loss or the original L_2^2 loss. Instead, we prove a weaker property that establishes strong correlation between the gradient of the empirical surrogate loss $\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) = (1/m) \sum_{i=1}^m (\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - y^{(i)})\mathbf{x}^{(i)}$ and the direction $\mathbf{w}^t - \mathbf{w}^*$ that holds whenever $\mathbf{w}^t$ is not an $O(\mathrm{OPT}) + \epsilon$ error solution and which is *independent of the scale* of $\mathbf{w}^t$. This constitutes our key structural result, stated as Proposition 3.1 and discussed in detail in Section 3. We further discuss how this result relates to classical and recent local error bounds in Appendix B.

In addition to this weaker version of a sharpness property, we further prove in Corollary 3.4 that given a parameter $\mathbf{w}^t$ and a dataset of m samples from $\mathcal{D}$, the activation $\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x})$ generated by optimizing the empirical risk on the dataset as in (P) satisfies $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \lesssim b^2 \|\mathbf{w}^t - \mathbf{w}^*\|_2^2$ with high probability. As a result, we can guarantee that when $\|\mathbf{w}^t - \mathbf{w}^*\|_2$ decreases, the L_2^2 distance between $\hat{u}^t$ and u^* diminishes as well. This is crucial, since without such a coupling we would not be able to argue about convergence over both model parameters $u, \mathbf{w}$.

Leveraging these results, we arrive at an algorithm that alternates between "gradient descent-style" updates for $\mathbf{w}$ and best-fit updates for u. We note in passing that similar alternating updates have been used in classical work on SIM learning in the less challenging, non-agnostic setting (Kakade et al., 2011). In more detail, our algorithm fixes the scale β of $\|\mathbf{w}^t\|_2$ and alternates between taking a Riemannian gradient descent step on a sphere for $\mathbf{w}^t$ w.r.t. the empirical surrogate loss and solving (P). The unknown scale for the true parameter vector $\mathbf{w}^*$ is resolved by applying this approach using β chosen from a sufficiently fine grid of the interval $[0, W]$ and employing a testing procedure at the end to select the best parameter vector. Although the idea is simple, the proof is quite technical, as it requires ensuring that the entire process does not accumulate spurious errors arising from the stochastic nature of the problem, adversarial labels, and approximate minimization of the surrogate loss, and, as a result, that it converges to the target error.

Technical Comparison to Gollakota et al. (2023) The only prior work addressing SIM learning (with unknown link functions) in the agnostic model is Gollakota et al. (2023), thus here we provide a technical comparison. While both Gollakota et al. (2023) and our work make use of the surrogate loss function from (2), on a technical level the two works are completely disjoint. Gollakota et al. (2023) uses a framework of omnipredictors to minimize the surrogate loss and then relates this result to the L_2^2 loss. Although they handle more general distributions and activations, their learner outputs a hypothesis with error that cannot be considered constant factor approximation (see (1)) and is improper. By contrast, our work does not seek to minimize the surrogate loss. Instead, our main insight is that the gradient of the surrogate loss at a vector $\mathbf{w}$ conveys information about the direction of a target vector $\mathbf{w}^*$, for a *fixed* link function that minimizes *the L_2^2 loss*. We leverage this property to construct a proper learner with constant factor approximation.

2. Preliminaries

Basic Notation For $n \in \mathbb{Z}_+$, let $[n] \coloneqq \{1, \ldots, n\}$. We use lowercase boldface characters for vectors. We use $\mathbf{x} \cdot \mathbf{y}$ for the inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\theta(\mathbf{x}, \mathbf{y})$ for the angle between $\mathbf{x}, \mathbf{y}$. For $\mathbf{x} \in \mathbb{R}^d$ and $k \in [d]$, $\mathbf{x}_k$ denotes the k^{th} coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2$ denotes the ℓ_2-norm of $\mathbf{x}$. We use $\mathbb{1}_A = \mathbb{1}\{A\}$ to denote the characteristic function of the set A. For vectors $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$, we denote by $\mathbf{v}^{\perp \mathbf{u}}$ the projection of $\mathbf{v}$ onto the subspace orthogonal to $\mathbf{u}$, i.e., $\mathbf{v}^{\perp \mathbf{u}} \coloneqq \mathbf{v} - ((\mathbf{v} \cdot \mathbf{u})\mathbf{u})/\|\mathbf{u}\|_2^2$. We use $\mathbb{B}(r)$ to denote the ℓ_2 ball in $\mathbb{R}^d$ of radius r, centered at the origin.

Asymptotic Notation We use the standard $O(\cdot), \Theta(\cdot), \Omega(\cdot)$ asymptotic notation. We use $\widetilde{O}(\cdot)$ to omit polylogarithmic factors in the argument. We use $O_p(\cdot)$ to suppress polynomial dependence on p, i.e., $O_p(\omega) = O(\mathrm{poly}(p)\omega)$. $\Theta_p(\cdot)$ and $\Omega_p(\cdot)$ are defined similarly. We write $E \gtrsim F$ for 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 $E \geq c\, F$. The notation $\lesssim$ is defined similarly.

Probability Notation We use $\mathbf{E}_{X \sim \mathcal{D}}[X]$ for the expectation of a random variable X according to the distribution $\mathcal{D}$ and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. For simplicity of notation, we omit the distribution when it is clear from the context. For $(\mathbf{x}, y)$ distributed according to $\mathcal{D}$, we use $\mathcal{D}_\mathbf{x}$ to denote the marginal distribution of $\mathbf{x}$.

Organization In Section 3, we establish our main structural result of alignment sharpness. In Section 4, we describe and analyze our constant factor approximate SIM learner. We conclude the paper in Section 5. The full version of the proofs is deferred to the supplementary material.

3. Main Structural Result: Alignment Sharpness of Surrogate Loss

In this section, we establish our main structural result (Proposition 3.1), which is what crucially enables us to obtain the target $O(\mathrm{OPT}) + \epsilon$ error for the studied problem.

Proposition 3.1 states that the empirical gradient of the surrogate loss positively correlates with the direction of $\mathbf{w}^t - \mathbf{w}^*$ whenever $\mathbf{w}^t$ does not correspond to an $O(\mathrm{OPT}) + \epsilon$ solution, and, moreover, the correlation is proportional to the quantity $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$. This is a key property that is leveraged in our algorithmic result (Theorem 4.2), both in obtaining an $O(\mathrm{OPT}) + \epsilon$ error result and in arguing about the convergence and computational efficiency of our algorithm. Intuitively, what Proposition 3.1 allows us to argue is that as long as the angle between $\mathbf{w}^t$ and $\mathbf{w}^*$ is not close to zero, we can update $\mathbf{w}^t$ to better align it with $\mathbf{w}^*$, in the sense that we reduce the angle between these two vectors.

Proposition 3.1 (Alignment Sharpness of the Convex Surrogate)**.** *Suppose that $\mathcal{D}_\mathbf{x}$ is (L, R)-well-behaved, $\mathcal{U}_{(a,b)}$ is as in Definition 1.3, and $\epsilon, \delta > 0$. Let $\mu \gtrsim a^2 L R^4/b$. Given any $\mathbf{w}^t \in \mathbb{B}(W)$, denote by $\hat{u}^t$ the optimal solution to (P) with respect to $\mathbf{w}^t$ and the sample set $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ drawn i.i.d. from $\mathcal{D}$. If m satisfies*

$$m \gtrsim dW^{9/2} b^4 L^{-4} \log^4(d/(\epsilon\delta))(1/\epsilon^{3/2} + 1/(\epsilon\delta)) \,,$$

then, with probability at least $1 - \delta$,

$$\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) \geq \mu \|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2 - 2(\mathrm{OPT} + \epsilon)/b - 2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2 \,.$$

To prove Proposition 3.1, we rely on the following key ingredients. In Section 3.1, we prove our main technical lemma (Lemma 3.2), which states that the L_2^2 distance between a hypothesis $u(\mathbf{w}\cdot\mathbf{x})$ and the target $u^*(\mathbf{w}^*\cdot\mathbf{x})$ is bounded below by the misalignment of $\mathbf{w}^t$ and $\mathbf{w}^*$, i.e., the squared norm of the component of $\mathbf{w}^*$ that is orthogonal to $\mathbf{w}^t$, $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$. As will become apparent in the proof of Proposition 3.1, the inner product $\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*)$ can be bounded below as a function of the empirical L_2^2 error for $\mathbf{w}^t$ and a different (but related) activation $\hat{u}^{*t}$, which can in turn be argued to be close to the population L_2^2 error for a sufficiently large sample size, using concentration. Thus, Lemma 3.2 can be leveraged to obtain a term scaling with $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$ in the lower bound on $\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*)$.

In Section 3.2, we characterize structural properties of the population-optimal link functions u^t and u^{*t} (see (EP) and (EP*)), which play a crucial role in the proof of Proposition 3.1. Specifically, we show that activation u^t is close to the idealized activation u^{*t} (the optimal activation without noise, given $\mathbf{w}^t$) in L_2^2 distance (Lemma 3.3). Since by standard uniform convergence results we have that $\hat{u}^t$ and $\hat{u}^{*t}$ are close to their population counterparts u^t and u^{*t}, respectively, Lemma 3.3 certifies that $\hat{u}^t$ is not far away from $\hat{u}^{*t}$. This property enables us to replace $\hat{u}^t$ by (the idealized) $\hat{u}^{*t}$ in the empirical surrogate gradient $\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)$, which is easier to analyze, since $\hat{u}^{*t}$ is defined with respect to the "ideal" dataset (with uncorrupted labels).

Finally, as a simple corollary of Lemma 3.3, we prove Corollary 3.4, which gives a clear explanation of why our algorithm, which alternates between updating $\mathbf{w}^t$ and $\hat{u}^t$, works: we show that the L_2^2 loss between the hypothesis generated by our algorithm $\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x})$ and the underlying optimal hypothesis $u^*(\mathbf{w}^* \cdot \mathbf{x})$ is bounded above by the distance between $\mathbf{w}^t$ and $\mathbf{w}^*$. Since our structural sharpness result (Proposition 3.1) enables us to decrease $\|\mathbf{w}^t - \mathbf{w}^*\|_2$, Corollary 3.4 certifies that choosing the empirically-optimal activation leads to convergence of the hypothesis $\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x})$.

Equipped with these technical lemmas, we prove our main structural result (Proposition 3.1) in Section 3.3.

3.1. L_2^2 Error and Misalignment

Our first key result is Lemma 3.2 below, which plays a critical role in the proof of Proposition 3.1. As discussed in Section 1.2, for two different activations u and u^* and parameters $\mathbf{w}$ and $\mathbf{w}^*$ such that $\mathbf{w}$ and $\mathbf{w}^*$ are parallel, even when the L_2^2 error is $\Omega(1)$, the gradient $\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}; u)$ might not significantly align with the direction of $\mathbf{w} - \mathbf{w}^*$, and thus cannot provide sufficient information about the direction to decrease $\|\mathbf{w} - \mathbf{w}^*\|_2$. Intuitively, the following lemma shows that this is the only thing that can go wrong, and it happens when $\mathbf{w}$ and $\mathbf{w}^*$ are parallel. In particular, Lemma 3.2 shows that for any square integrable link function f, we can relate the L_2^2 distance $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]$ to the magnitude of the component of $\mathbf{w}^*$ that is *orthogonal* to $\mathbf{w}$. The full proof of Lemma 3.2 is deferred to Appendix C.2.

Lemma 3.2 (Lower-Bounding L_2^2 Error by Misalignment)**.** *Let $u^* \in \mathcal{U}_{(a,b)}$, $\mathcal{D}_\mathbf{x}$ be (L, R)-well-behaved, and $f : \mathbb{R} \mapsto \mathbb{R}$ be square-integrable with respect to the measure of the distribution $\mathcal{D}_\mathbf{x}$. Then, for any $\mathbf{w}, \mathbf{w}^* \in \mathbb{R}^d$, it holds that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \gtrsim a^2 L R^4 \|(\mathbf{w}^*)^{\perp_\mathbf{w}}\|_2^2 \,.$$

Proof Sketch of Lemma 3.2. Suppose for simplicity that $R = 1$, and let us denote $(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2$ by $p(\mathbf{x})$. The statement holds trivially when $\mathbf{w}$ is parallel to $\mathbf{w}^*$, so assume this is not the case. Let us also assume $\mathbf{w}^*\cdot\mathbf{w} \geq 0$, as for the case of $\mathbf{w}^*\cdot\mathbf{w} \leq 0$ the proof can be carried out using similar arguments. Define $\mathbf{v} = (\mathbf{w}^*)^{\perp_\mathbf{w}} = \mathbf{w}^* - (\mathbf{w}^*\cdot\mathbf{w})\mathbf{w}/\|\mathbf{w}\|_2^2$ and let $\tilde{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_2$. Then $\mathbf{w}^* = \alpha\mathbf{w} + \mathbf{v}$, for some $\alpha \geq 0$. Let V be the subspace spanned by $\mathbf{w}, \mathbf{v}$. Then considering the event $A = \{\mathbf{w}\cdot\mathbf{x} \geq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (1/16, 1/8)\cup(3/8, 1/2)\}$, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{x})] \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{w}\cdot\mathbf{x}_V) - u^*(\mathbf{w}^*\cdot\mathbf{x}_V))^2 \mathbb{1}\{A\}]$.

The main idea is to study the relation between the value of $f(\mathbf{w}\cdot\mathbf{x})$ and $u^*(\mathbf{w}^*\cdot\mathbf{x})$, utilizing the fact that $u^*(z)$ is strictly increasing when $z \geq 0$. In particular, define the following functions indicating the intervals that $f(\mathbf{w}\cdot\mathbf{x})$ belongs to: $I_1(\mathbf{x}) = \mathrm{sign}(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2/32))$,

$I_2(\mathbf{x}) = \mathrm{sign}(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2/4))$, and $I_3(\mathbf{x}) = \mathrm{sign}(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2))$. Since u^* is non-decreasing, it must be $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$ or $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$. We discuss the cases of $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$ and $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$, and provide lower bound for each case respectively.

Consider first $I_1(\mathbf{x}), I_2(\mathbf{x}) \geq 0$, indicating that $f(\mathbf{w}\cdot\mathbf{x}) \geq u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2/4)$. Further restricting $\mathbf{x}$ on the band $B = \{\mathbf{w}\cdot\mathbf{x} \geq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (1/16, 1/8)\}$, $B \subset A$, we have $u^*(\mathbf{w}^*\cdot\mathbf{x}) \leq u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2/8)$. Using the fact that $u^*(z) - u^*(z') \geq a(z - z')$ when $z \geq z' \geq 0$, we have

$$\begin{aligned} p(\mathbf{x})\mathbb{1}\{B\} &= \big(\{f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2/4)\} \\ &\quad + \{u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2/4) - u^*(\mathbf{w}^*\cdot\mathbf{x})\}\big)^2\mathbb{1}\{B\} \\ &\geq (u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2/4) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{B\} \\ &\geq (a^2/8^2)\|\mathbf{v}\|_2^2\mathbb{1}\{B\}\,. \end{aligned}$$

With similar arguments, when both $I_1(\mathbf{x}), I_2(\mathbf{x}) \leq 0$, it holds $p(\mathbf{x})\mathbb{1}\{B\} \geq (a^2/32)^2\|\mathbf{v}\|_2^2\mathbb{1}\{B\}$. Thus, when $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$ we have $p(\mathbf{x})\mathbb{1}\{B\} \gtrsim a^2\|\mathbf{v}\|_2^2\mathbb{1}\{B\}$.

For the case of $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$, we consider restricting $\mathbf{x}$ on $B' = \{\mathbf{w}\cdot\mathbf{x} \geq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (3/8, 1/2)\}$. Then, similarly, after discussing the cases of $I_2(\mathbf{x}), I_3(\mathbf{x}) \geq 0$ and $I_2(\mathbf{x}), I_3(\mathbf{x}) \leq 0$, we get that when $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$, $p(\mathbf{x})\mathbb{1}\{B'\} \gtrsim a^2\|\mathbf{v}\|_2^2\mathbb{1}\{B'\}$.

Finally, let $D = (B \cap \{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\}) \cup (B' \cap \{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}) \subset A$. Since $\mathcal{D}_\mathbf{x}$ is $(L, 1)$-well behaved, the probability mass γ_V satisfies $\gamma_V(\mathbf{x}) \geq L$ when $\|\mathbf{x}\|_\infty \leq 1$. Therefore, since V is a 2-dimensional plane, using geometric observations, $\mathbf{Pr}[D]$ can be bounded below by $\mathbf{Pr}[D] \geq \mathbf{Pr}[D \cap \{\|\mathbf{x}\|_\infty \leq 1\}] \geq L/16$. Hence, combining the bounds above and recalling that $b \geq a$, we finally get $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{x})] \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{x})\mathbb{1}\{D\}] \gtrsim a^2\|\mathbf{v}\|_2^2\,\mathbf{Pr}[D] \gtrsim a^2L\|\mathbf{v}\|_2^2$ completing the proof of Lemma 3.2. □

3.2. Closeness of Idealized and Attainable Activations

In this section, we bound the contribution of the error incurred from working with attainable link functions $\hat{u}^t$ in the iterations of the algorithm. The error incurred is due to both the arbitrary noise in the labels and due to using a finite sample set. In bounding the error, for analysis purposes, we introduce auxiliary population-level link functions.

Concretely, given $\mathbf{w} \in \mathbb{B}(W)$, a population-optimal activation is a solution to the following stochastic convex program:

$$u_\mathbf{w} \in \operatorname*{argmin}_{u\in\mathcal{U}_{(a,b)}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}\cdot\mathbf{x}) - y)^2]. \tag{EP}$$

We further introduce auxiliary "idealized, noiseless" activations, which, given noiseless labels $y^* = u^*(\mathbf{w}^*\cdot\mathbf{x})$ and a parameter weight vector $\mathbf{w}$, are defined via

$$u^*_\mathbf{w} \in \operatorname*{argmin}_{u\in\mathcal{U}_{(a,b)}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}\cdot\mathbf{x}) - y^*)^2]. \tag{EP*}$$

Below we relate $u^t := u_{\mathbf{w}^t}$ and $u^{*t} := u^*_{\mathbf{w}_t}$ and show that their L_2^2 error for the parameter vector $\mathbf{w}^t$ is bounded by OPT. The proof of Lemma 3.3 is deferred to Appendix C.3.

Lemma 3.3 (Closeness of Population-Optimal Activations)**.** *Let $\mathbf{w}^t \in \mathbb{B}(W)$ and let u^{*t}, u^t be defined as solutions to (EP*), (EP), respectively. Then,*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] \leq \mathrm{OPT}.$$

As a consequence of the lemma above, we are able to relate $\hat{u}^t$ to the "noiseless" labels $y^* = u^*(\mathbf{w}^*\cdot\mathbf{x})$ by showing that the L_2^2 distance between $u^*(\mathbf{w}^*\cdot\mathbf{x})$ and the sample-optimal activation $\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})$ is bounded by $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2$. Although Corollary 3.4 is not used in the proof of Proposition 3.1, we still present it here as it justifies the mechanism of our approach alternating between updates for $\mathbf{w}^t$ and $\hat{u}^t$. The proof of Corollary 3.4 can be found in Appendix C.4.

Corollary 3.4 (Closeness of Idealized and Attainable Activations)**.** *Let $\epsilon, \delta > 0$. Given a parameter $\mathbf{w}^t \in \mathbb{B}(W)$ and $m \gtrsim d\log^4(d/(\epsilon\delta))(b^2W^3/(L^2\epsilon))^{3/2}$ samples from $\mathcal{D}$, let $\hat{u}^t$ be the sample-optimal activation on these samples given $\mathbf{w}^t$, as defined in (P). Then, with probability at least $1 - \delta$,*

$$\begin{aligned} &\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \\ &\qquad \leq 3(\epsilon + \mathrm{OPT} + b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2). \end{aligned}$$

3.3. Proof of Proposition 3.1

We now provide a proof sketch for Proposition 3.1, while the detailed proof is deferred to Appendix C.1.

Proof Sketch of Proposition 3.1. Given any weight parameter $\mathbf{w}^t \in \mathbb{B}(W)$ and $\hat{u}^t$ as defined in the statement, let u^t be the population-optimal activation defined by (EP). Given $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$, define $y^{*(i)} = u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$ for $i \in [m]$. Further define the following auxiliary problem:

$$\hat{u}^*_\mathbf{w} \in \operatorname*{argmin}_{u\in\mathcal{U}_{(a,b)}} \frac{1}{m}\sum_{i=1}^m (u(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{*(i)})^2. \tag{P*}$$

Denote $\hat{u}^{*t} := \hat{u}^*_{\mathbf{w}^t}$. Observe that (P*) is the empirical version of (EP*). To prove Proposition 3.1, we decompose

$\nabla\widehat{\mathcal{L}}_{\rm sur}(\mathbf{w}^t;\hat{u}^t)\cdot(\mathbf{w}^t-\mathbf{w}^*)$ into three summation terms.

$$\begin{aligned}
&\nabla\widehat{\mathcal{L}}_{\rm sur}(\mathbf{w}^t;\hat{u}^t)\cdot(\mathbf{w}^t-\mathbf{w}^*)\\
&=\underbrace{\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}}_{Q_1}\\
&\quad+\underbrace{\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}}_{Q_2}\\
&\quad+\underbrace{\frac{1}{m}\sum_{i=1}^{m}(y^{*(i)}-y^{(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\mathbf{w}^*\cdot\mathbf{x}^{(i)})}_{Q_3}. \qquad (4)
\end{aligned}$$

We bound each term Q_1, Q_2, Q_3 in (4) separately and defer the proofs of corresponding claims to Appendix C. We first show that with probability at least $1-\delta$, we have

$$Q_1 \geq -(\sqrt{\epsilon}+\sqrt{\mathrm{OPT}})\|\mathbf{w}^t-\mathbf{w}^*\|_2-(\mathrm{OPT}+\epsilon)/b. \quad (5)$$

Inequality (5) contains two error terms, the first one is due to noise and the second one is the estimation error. The second term comes from standard concentration results (see Appendix F). The first term comes from replacing $\hat{u}^t$ and $\hat{u}^{*t}$ with their population counterparts u^t and u^{*t} and combining Cauchy-Schwarz inequality with Lemma 3.3.

We next show that Q_2 is a constant multiple of $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$, which is the main positive contribution among the three summands. In particular, we show that for an absolute constant C', with probability at least $1-\delta$,

$$Q_2 \geq \frac{C'a^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2-\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2-\epsilon/b. \quad (6)$$

The proof of the above statement is rather technical. We first define an 'empirical inverse' $\hat{f}:\mathbb{R}\to\mathbb{R}$ of the activation u^*, as u^* is not strictly increasing hence $(u^*)^{-1}$ is not well-defined on $\mathbb{R}$. Denote $q(\mathbf{x}^{(i)}):=\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$. Then adding and subtracting $\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))$,

$$\begin{aligned}
Q_2 &= \frac{1}{m}\sum_{i=1}^{m}q(\mathbf{x}^{(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})))\\
&\quad+\frac{1}{m}\sum_{i=1}^{m}q(\mathbf{x}^{(i)})(\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)}).
\end{aligned}$$

Using the optimality conditions of (P*), the first summation above is always positive. For the second summation, by the definition of $\hat{f}$, we have that $|\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)}|\geq(1/b)|q(\mathbf{x}^{(i)})|$ and $q(\mathbf{x}^{(i)})(\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)})\geq 0$, leading to a lower bound of $(1/bm)\sum_{i=1}^{m}(q(\mathbf{x}^{(i)}))^2$. Using Lemma 3.2 and standard concentration arguments, we then obtain Inequality (6). Finally, using similar arguments as for Q_1, we show that with probability at least $1-\delta$,

$$Q_3 \geq -\sqrt{\mathrm{OPT}}\|\mathbf{w}^*-\mathbf{w}^t\|_2-(\mathrm{OPT}+\epsilon)/b\,. \quad (7)$$

The proof is completed by plugging the bounds from Inequalities (5) to (7) back into Equation (4). □

4. Robust SIM Learning via Alignment Sharpness

As discussed in Section 1.2, our algorithm can be viewed as employing an alternating procedure: taking a Riemannian gradient descent step on a sphere with respect to the empirical surrogate, given an estimate of the activation, and optimizing the activation function on the sample set for a given parameter weight vector. This procedure is performed using a fine grid of guesses of the scale of $\|\mathbf{w}^*\|_2$. For this process to converge with the desired linear rate (even for a known value of $\|\mathbf{w}^*\|_2$), the algorithm needs to be properly initialized to ensure that the initial weight vector has a nontrivial alignment with the optimal vector $\mathbf{w}^*$. The initialization process is handled in the following subsection.

4.1. Initialization

We begin by showing that the Initialization subroutine stated in Algorithm 2 (see Appendix D.1) returns a point $\bar{\mathbf{w}}^0$ that has a sufficient alignment with $\mathbf{w}^*$. As will become apparent later in the proof of Theorem 4.2, this property of the initial point is critical for Algorithm 1 to converge at a linear rate. We defer the more detailed statement of Lemma 4.1 (see Lemma D.1) and its proof to Appendix D.1.

Lemma 4.1 (Initialization). *Given $\mu=O_{L,R}(a^2/b)$ and $\epsilon,\delta>0$, (Initialization) Algorithm 2 draws $m_0=\tilde{O}_{W,b,1/L,1/\mu}(d/(\delta\epsilon^{3/2}))$ i.i.d. samples from $\mathcal{D}$, it runs in time $\tilde{O}_{W,b,1/L,1/\mu}(dm_0)$, and with probability at least $1-\delta$, it generates a list of size $t_0\lesssim(b/\mu)^6\log(b/\mu)$ that contains a vector $\bar{\mathbf{w}}^0$ such that*

$$\|(\mathbf{w}^*)^{\perp_{\bar{\mathbf{w}}^0}}\|_2\leq\max\{\mu\|\mathbf{w}^*\|_2/b,\, b^2/\mu^3(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})\}.$$

4.2. Optimization

Our main optimization algorithm is summarized in Algorithm 1 (see Algorithm 3 for a more detailed version). We now provide intuition for how guessing the value of $\|\mathbf{w}^*\|_2$ is used in the convergence analysis. Let $\mathbf{w}^t=\|\mathbf{w}^*\|_2\bar{\mathbf{w}}^t/\|\bar{\mathbf{w}}^t\|_2$ so that $\|\mathbf{w}^t\|_2=\|\mathbf{w}^*\|_2$ and let $\mathbf{v}^t:=(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}$. Observe that $\|\mathbf{v}^t\|_2=\|\mathbf{w}^t-\mathbf{w}^*\|_2\cos(\theta(\mathbf{w}^t,\mathbf{w}^*)/2)$. Applying Proposition 3.1, it can be shown that $\|\bar{\mathbf{w}}^{t+1}-\mathbf{w}^*\|_2^2\leq\|\mathbf{w}^t-\mathbf{w}^*\|_2^2-C\|\mathbf{v}^t\|_2^2$ for some constant C. Thus, as long as the angle between $\mathbf{w}^t$ and $\mathbf{w}^*$ is not too large (ensured by initialization), $\|\mathbf{w}^t-\mathbf{w}^*\|_2\approx\|\mathbf{v}^t\|_2$. Hence, we can argue that

$\|\mathbf{w}^t - \mathbf{w}^*\|_2$ contracts in each iteration, by observing that $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \approx \|\mathbf{v}^{t+1}\|_2^2 \le \|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2$.

Algorithm 1 Optimization

Input: $\mathbf{w}^{\mathrm{ini}} = \mathbf{0}$; $\epsilon > 0$; positive parameters: a, b, L, R, W, μ; step size η
$\{\mathbf{w}_0^{\mathrm{ini}}, \dots, \mathbf{w}_{t_0}^{\mathrm{ini}}\} = \text{Initialization}[\mathbf{w}^{\mathrm{ini}}]$ (Algorithm 2)
$\mathcal{P} = \{(\mathbf{w} = 0; u(z) = 0)\}$
for $k = 0$ **to** $t_0 \lesssim (b/\mu)^6 \log(b/\mu)$ **do**
  for $j = 1$ **to** $J = W/(\eta\sqrt{\epsilon})$ **do**
    $\bar{\mathbf{w}}_{j,k}^0 = \mathbf{w}_k^{\mathrm{ini}}$, $\beta_j = j\eta\sqrt{\epsilon}$
    for $t = 0$ **to** $T = O((b/\mu)^2 \log(1/\epsilon))$ **do**
      $\widehat{\mathbf{w}}_{j,k}^t = \beta_j(\bar{\mathbf{w}}_{j,k}^t / \|\bar{\mathbf{w}}_{j,k}^t\|_2)$
      Draw $m = \tilde{\Theta}_{W,b,1/L,1/\mu}(d/\epsilon^{3/2})$ new samples
      $\hat{u}_{j,k}^t = \operatorname{argmin}_{u \in \mathcal{U}_{(a,b)}} \frac{1}{m} \sum_{i=1}^m (u(\widehat{\mathbf{w}}_{j,k}^t \cdot \mathbf{x}^{(i)}) - y^{(i)})^2$
      $\bar{\mathbf{w}}_{j,k}^{t+1} = \widehat{\mathbf{w}}_{j,k}^t - \eta \nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}_{j,k}^t; \hat{u}_{j,k}^t)$
    end for
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{(\widehat{\mathbf{w}}_{j,k}^T; \hat{u}_{j,k}^T)\}$
  end for
end for
$(\widehat{\mathbf{w}}; \hat{u}) = \text{Test}[(\mathbf{w}; u) \in \mathcal{P}]$ (Algorithm 4)
Return: $(\widehat{\mathbf{w}}; \hat{u})$

Our main result is the following theorem (see Theorem D.2 for a more detailed statement and proof in Appendix D.2):

Theorem 4.2 (Main Result)**.** *Let $\mathcal{D}$ be a distribution in $\mathbb{R}^d \times \mathbb{R}$ and suppose that $\mathcal{D}_{\mathbf{x}}$ is (L, R)-well-behaved. Let $\mathcal{U}_{(a,b)}$ be as in Definition 1.3 and let $\epsilon > 0$. Then, Algorithm 1 uses $N = \tilde{O}_{W,b,1/L,1/\mu}(d/\epsilon^2)$ samples, it runs for $\tilde{O}_{W,b,1/\mu}(1/\sqrt{\epsilon})$ iterations, and, with probability at least $2/3$, returns a hypothesis $(\hat{u}, \widehat{\mathbf{w}})$, where $\hat{u} \in \mathcal{U}_{(a,b)}$ and $\widehat{\mathbf{w}} \in \mathbb{B}(W)$, such that $\mathcal{L}_2(\widehat{\mathbf{w}}; \hat{u}) = O_{1/L,1/R,b/a}(\mathrm{OPT}) + \epsilon$.*

To prove Theorem 4.2, we make use of two technical results stated below. First, Lemma 4.3 provides an upper bound on the norm of the empirical gradient of the surrogate loss. The proof of the lemma relies on concentration properties of (L, R)-well behaved distributions $\mathcal{D}_{\mathbf{x}}$, and leverages the uniform convergence of the empirically-optimal activations $\hat{u}^t$. A more detailed statement (Lemma D.7) and the proof of Lemma 4.3 is deferred to Appendix D.3.

Lemma 4.3 (Bound on Empirical Gradient Norm)**.** *Let S be a set of i.i.d. samples of size $m = \tilde{\Theta}_{W,b,1/L}(d/\epsilon^{3/2} + d/(\epsilon\delta))$. Given any $\mathbf{w}^t \in \mathbb{B}(W)$, let $\hat{u}^t \in \mathcal{U}_{(a,b)}$ be the solution of optimization problem (P) with respect to $\mathbf{w}^t$ and sample set S. Then, with probability at least $1 - \delta$,*

$$\|\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2^2 \le 4b^2 \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 10(\mathrm{OPT} + \epsilon).$$

The next claim bounds the L_2^2 error of a hypothesis $\hat{u}_{\mathbf{w}}(\mathbf{w} \cdot \mathbf{x})$ by the distance between $\mathbf{w}$ and $\mathbf{w}^*$. We defer a more detailed statement (Claim D.8) and the proof to Appendix D.4.

Claim 4.4. Let $\mathbf{w} \in \mathbb{B}(W)$ be any fixed vector. Let $\hat{u}_{\mathbf{w}}$ be defined by (P) given $\mathbf{w}$ and a sample set of size $m = \tilde{\Theta}_{W,b,1/L}(d/\epsilon^{3/2})$. Then, $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}_{\mathbf{w}}(\mathbf{w} \cdot \mathbf{x}) - y)^2] \le 8(\mathrm{OPT} + \epsilon) + 4b^2\|\mathbf{w} - \mathbf{w}^*\|_2^2$.

Proof Sketch of Theorem 4.2. For this sketch, we consider the case $\|\mathbf{w}^*\|_2 \gtrsim b^3/\mu^4(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$ so that the initialization subroutine generates a point $\mathbf{w}_{k^*}^{\mathrm{ini}} \in \{\mathbf{w}_k^{\mathrm{ini}}\}_{k=1}^{t_0}$ such that $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}_{k^*}^{\mathrm{ini}}}}\|_2 \le \mu\|\mathbf{w}^*\|_2/(4b)$, by Lemma 4.1. Fix this initialized parameter $\bar{\mathbf{w}}_{j,k^*}^0 = \mathbf{w}_{k^*}^{\mathrm{ini}}$ at step k^* and drop the subscript k^* for simplicity. Since we constructed a grid with width $\eta\sqrt{\epsilon}$, there exits an index j^* such that $|\beta_{j^*} - \|\mathbf{w}^*\|_2| \le \eta\sqrt{\epsilon}$. We consider the intermediate for-loop at this iteration j^*, and show that the inner loop with normalizer β_{j^*} outputs a solution with error $O(\mathrm{OPT}) + \epsilon$. This solution can be picked using standard testing procedures. We now focus on the iteration j^*, and drop the subscript j^* for notational simplicity.

Let $\mathbf{w}^t = \|\mathbf{w}^*\|_2(\bar{\mathbf{w}}^t/\|\bar{\mathbf{w}}^t\|_2)$ and denote $\mathbf{v}^t := (\mathbf{w}^*)^{\perp_{\widehat{\mathbf{w}}^t}}$. Expanding $\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2$ and applying Proposition 3.1 and Lemma 4.3, we get

$$\begin{aligned}
&\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 = \|\widehat{\mathbf{w}}^t - \eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t) - \mathbf{w}^*\|_2^2 \\
&= \|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2 + \eta^2\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t)\|_2^2 \\
&\quad - 2\eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t) \cdot (\widehat{\mathbf{w}}^t - \mathbf{w}^*) \\
&\le \|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2 + \eta^2(10(\mathrm{OPT} + \epsilon) + 4b^2\|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2) \\
&\quad + 2\eta(2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2 - \mu\|\mathbf{v}^t\|_2^2) \\
&\quad + 4\eta(\mathrm{OPT} + \epsilon)/b \\
&\le (1 + 4\eta^2b^2)\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + (24\eta^2 + 4\eta/b)(\mathrm{OPT} + \epsilon) \\
&\quad + 2\eta(2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2 - \mu\|\mathbf{v}^t\|_2^2), \qquad (8)
\end{aligned}$$

where in the last inequality we used $\|\widehat{\mathbf{w}}^t - \mathbf{w}^t\|_2 \le \eta\sqrt{\epsilon}$.

Since $\mathbf{w}^t$ and $\mathbf{w}^*$ are on the same sphere, $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \le \|\mathbf{v}^t\|_2$. In particular, letting $\rho_t = \|\mathbf{v}^t\|_2/\|\mathbf{w}^*\|_2$, we have $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \le (1 + \rho_t^2)\|\mathbf{v}^t\|_2^2 \le 2\|\mathbf{v}^t\|_2^2$. Recall that the algorithm is started from $\bar{\mathbf{w}}^0$ that satisfies $\rho_0 \le \mu/(4b)$. If $\rho_t \le \mu/(4b)$, then $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \le (1 + (\mu/(4b))^2)\|\mathbf{v}^t\|_2^2$. Assuming in addition that $\|\mathbf{v}^t\|_2 \gtrsim (1/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, and choosing the stepsize $\eta = \mu/(4b^2)$, (8) implies that

$$\|\mathbf{v}^{t+1}\|_2^2 \le \|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \le (1 - \mu^2/(32b^2))\|\mathbf{v}^t\|_2^2,$$

and thus, in addition, $\rho_{t+1} \le \mu/(4b)$. Therefore, by an inductive argument, we show that as long as $\mathbf{w}^t$ is still far from $\mathbf{w}^*$, i.e., $\|\mathbf{v}^t\|_2 \gtrsim (1/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, we have

$$\|\mathbf{v}^{t+1}\|_2^2 \le (1 - \mu^2/(32b^2))\|\mathbf{v}^t\|_2^2 \text{ and } \rho_{t+1} \le \mu/(4b).$$

Hence, after $T = O((b^2/\mu^2)\log(1/\epsilon))$ iterations, it must be $\|\mathbf{v}^T\|_2 \lesssim (1/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, which implies

$$\|\mathbf{w}^T - \mathbf{w}^*\|_2^2 \le 2\|\mathbf{v}^T\|_2^2 = O(\mathrm{OPT}) + \epsilon.$$

Finally, by Claim 4.4, hypothesis $\hat{u}^T(\widehat{\mathbf{w}}^T \cdot \mathbf{x})$ achieves L_2^2-error $O(\mathrm{OPT}) + \epsilon$, which completes the proof. □

5. Conclusion

We presented the first constant-factor approximate SIM learner in the agnostic model, for the class of (a, b)-unbounded link functions under mild distributional assumptions. Immediate questions for future research involve extending these results to other classes of link functions. More specifically, our results require that b/a is bounded by a constant. It is an open question whether the constant-factor approximation result in the agnostic model can be extended to all b-Lipschitz functions (with $a = 0$). This question is open even when the link function is known to the learner.

Acknowledgements

NZ was supported in part by NSF Medium Award CCF-2107079 and NSF Award DMS-2023239. PW was supported in part by NSF Award DMS-2023239. ID was supported by NSF Medium Award CCF-2107079 and a DARPA Learning with Less Labels (LwLL) grant. JD was supported by the U. S. Office of Naval Research under award number N00014-22-1-2348.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Awasthi, P., Tang, A., and Vijayaraghavan, A. Agnostic learning of general ReLU activation using gradient descent. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29, 2016.

Bolte, J., Nguyen, T. P., Peypouquet, J., and Suter, B. W. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

Dalalyan, A. S., Juditsky, A., and Spokoiny, V. A new algorithm for estimating the effective dimension-reduction subspace. *The Journal of Machine Learning Research*, 9: 1647–1678, 2008.

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

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

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

Diakonikolas, I., Kane, D. M., Pittas, T., and Zarifis, N. 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*, 2021.

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

Diakonikolas, I., Kane, D. M., Kontonis, V., Tzamos, C., and Zarifis, N. Learning general halfspaces with general Massart noise under the Gaussian distribution. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 874–885. ACM, 2022b.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning a single neuron with adversarial label noise via gradient descent. In *Conference on Learning Theory (COLT)*, pp. 4313–4361, 2022c.

Diakonikolas, I., Kane, D. M., and Ren, L. Near-optimal cryptographic hardness of agnostically learning halfspaces and ReLU regression under Gaussian marginals. In *ICML*, 2023.

Dudeja, R. and Hsu, D. Learning single-index models in Gaussian space. In *Conference on Learning Theory, COLT*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1887–1930. PMLR, 2018.

Facchinei, F. and Pang, J.-S. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

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

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

Gollakota, A., Gopalan, P., Klivans, A. R., and Stavropoulos, K. Agnostically learning single-index models using omnipredictors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., et al. *Nonparametric and semiparametric models*, volume 1. Springer, 2004.

Haussler, D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

Hoffman, A. J. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49:263–265, 1952.

Hristache, M., Juditsky, A., and Spokoiny, V. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pp. 595–623, 2001.

Ichimura, H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of econometrics*, 58(1-2):71–120, 1993.

Jin, C., Ge, R., Netrapalli, P., Kakade, S., and Jordan, M. How to escape saddle points efficiently. In *International conference on machine learning*, pp. 1724–1732. PMLR, 2017.

Kakade, S. M., Kanade, V., Shamir, O., and Kalai, A. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

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

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811, 2016.

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

Liu, J., Cui, Y., and Pang, J.-S. Solving nonsmooth and nonconvex compound stochastic programs with applications to risk measure minimization. *Mathematics of Operations Research*, 2022.

Łojasiewicz, S. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

Łojasiewicz, S. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pp. 1575–1595, 1993.

Lu, C. and Hochbaum, D. S. A unified approach for a 1D generalized total variation problem. *Mathematical Programming*, 194(1-2):415–442, 2022.

Manurangsi, P. and Reichman, D. The computational complexity of training ReLU(s). *arXiv preprint arXiv:1810.04207*, 2018.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Roulet, V. and d'Aspremont, A. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.

Síma, J. Training a single sigmoidal neuron is hard. *Neural Computation*, 14(11):2709–2728, 2002.

Wang, P., Zarifis, N., Diakonikolas, I., and Diakonikolas, J. Robustly learning a single neuron via sharpness. *40th International Conference on Machine Learning*, 2023.

Zhang, H. and Yin, W. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

Zheng, Q. and Lafferty, J. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Supplemental Material

Organization The appendix is organized as follows. In Appendix A, we highlight some useful properties about the distribution class and the activation class. Appendix B provides a detailed introduction about the notion of local error bounds and its relation to our alignment sharpness structural result. In Appendix C, we provide detailed proofs that are omitted in Section 3, and in Appendix D we complete the proofs omitted in Section 4. In Appendix E, we provide a detailed discussion about computing the optimal empirical activation. Finally, in Appendix F we state and prove the standard uniform convergence results that are used throughout the paper.

A. Remarks about the Distribution Class and the Activation Class

In this section, we show that without the loss of generality we can assume that the parameters L, R in the distributional assumptions (Definition 1.2) can be taken less than 1, while the parameters a, b of the activations functions (see Definition 1.3) can be taken as $a, 1/b \leq 1$.

Remark A.1 (Distribution/Activation Parameters, (Definition 1.2 & Definition 1.3)). We observe that if a distribution $\mathcal{D}_{\mathbf{x}}$ is (L, R)-well-behaved, then it is also (L', R')-well-behaved for any $0 < L' \leq L, 0 < R' \leq R$. Hence, it is without loss of generality to assume that $L, R \in (0, 1]$. Similarly, if an activation is (a, b)-unbounded, it is also an (a, b')-unbounded activation with $b' \geq b$. Thus, we assume that $b \geq 1$. We can similarly assume $a \leq 1$.

In addition, we remark that the (L, R)-well behaved distributions are sub-exponential.

Remark A.2 (Sub-exponential Tails of Well-Behaved Distributions, Definition 1.2). Definition 1.2 might seem abstract, but to put it plain it implies that the random variable $\mathbf{x}$ has a $1/L$-sub-exponential tail, and that the pdf of the projected random variable $\mathbf{x}_V$ onto the space V is lower bounded by L. To see the first statement, given any unit vector $\mathbf{p}$, let $\mathbf{x}_{\mathbf{p}}$ be the projection of $\mathbf{x}$ onto the one-dimensional linear space $V_{\mathbf{p}} = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{z} = t\mathbf{p}, t \in \mathbb{R}\}$, i.e., $\mathbf{x}_{\mathbf{p}} = \mathbf{p} \cdot \mathbf{x} \in V_{\mathbf{p}}$. Then, by the anti-concentration and concentration property, we have

$$\mathbf{Pr}[|\mathbf{p} \cdot \mathbf{x}| \geq r] = \mathbf{Pr}[|\mathbf{x}_{\mathbf{p}}| \geq r] \leq \int_{|x| \geq r} \gamma(x)\,\mathrm{d}x \leq 2\int_r^\infty \frac{1}{L}\exp(-Lx)\,\mathrm{d}x = \frac{2}{L^2}\exp(-Lr),$$

which implies that $\mathbf{x}$ possess a sub-exponential tail.

B. Local Error Bounds and Alignment Sharpness

Given a generic optimization problem $\min_{\mathbf{w}} f(\mathbf{w})$ and a non-negative residual function $r(\mathbf{w})$ measuring the approximation error of the optimization problem, we say that the problem satisfies a local error bound if in some neighborhood of "test" (typically optimal) solutions $\mathcal{W}^*$ we have that

$$r(\mathbf{w}) \geq (\mu/\nu)\,\mathrm{dist}(\mathbf{w}, \mathcal{W}^*)^\nu. \tag{9}$$

In other words, low value of the residual function implies that $\mathbf{w}$ must be close to the test set $\mathcal{W}^*$.

Local error bounds have been studied in the optimization literature for decades, starting with the seminal works of (Hoffman, 1952; Łojasiewicz, 1963); see, e.g., Chapter 6 in (Facchinei & Pang, 2003) for an overview of classical results and (Bolte et al., 2017; Karimi et al., 2016; Roulet & d'Aspremont, 2017; Mei et al., 2018; Liu et al., 2022) and references therein for a more cotemporary overview. While local error bounds can be shown to hold generically under fairly minimal assumptions on f and for $r(\mathbf{w}) = f(\mathbf{w}) - \min_{\mathbf{w}'} f(\mathbf{w}')$ (Łojasiewicz, 1963; 1993), it is rarely the case that it can be ensured with a parameter μ that is not trivially small.

On the other hand, learning problems often possess very strong structural properties that can lead to stronger local error bounds. There are two main such examples we are aware of, where local error bounds can be shown to hold with $\nu = 2$ and an absolute constant $\mu > 0$. The first example are low-rank matrix problems such as matrix completion and matrix sensing, which are unrelated to our work (Bhojanapalli et al., 2016; Zheng & Lafferty, 2016; Jin et al., 2017). More relevant to our work are the recent results in (Mei et al., 2018; Wang et al., 2023) which proved local error bounds of the form

$$r(\mathbf{w}) \geq \frac{\mu}{2}\mathrm{dist}(\mathbf{w}, \mathcal{W}^*)^2 \tag{10}$$

for the more restricted problem than ours: (Mei et al., 2018) only dealt with the additive zero-mean noise, and was given the knowledge of the activation, and in addition, they assumed that the marginal $\mathcal{D}_{\mathbf{x}}$ is sub-Gaussian; while in (Wang et al.,

2023) they considered the agnostic learning of SIMs also with a known activation function but under somewhat more general distributional assumptions. In (Wang et al., 2023), the residual function was defined by $r(\mathbf{w}^t) = \nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; u^*)\cdot(\mathbf{w}^t - \mathbf{w}^*)$, where $\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; u^*)$ is the gradient of an empirical surrogate loss, and the resulting local error bound referred to as "sharpness."[2]

Our structural result can be seen as a weak notion of a local error bound, where the residual function for the empirical surrogate loss expressed as $r(\mathbf{w}^t, \hat{u}^t) = \nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\cdot(\mathbf{w}^t - \mathbf{w}^*)$ is bounded below as a function of the magnitude of the component of $\mathbf{w}^*$ that is orthogonal to $\mathbf{w}^t$. Compared to more traditional local error bounds and the bound from (Wang et al., 2023), which bound below the residual error function as a function of the distance to $\mathcal{W}^*$, this is a much weaker local error bound since it does not distinguish between vectors of varying magnitudes along the direction of $\mathbf{w}^*$. Since our lower bound is related to the "sharpness" notion studied in (Wang et al., 2023), we refer to it as the "alignment sharpness" to emphasize that it only relates the misalignment (as opposed to the distance) of vectors $\mathbf{w}^t$ and $\mathbf{w}^*$ to the residual error. To the best of our knowledge, such a form of a local error bound, which only bounds the alignment of vectors as opposed to their distance, is novel. We expect it to find a more broader use in learning theory and optimization.

C. Omitted Proofs from Section 3

C.1. Proof of Proposition 3.1

This subsection is devoted to the full version of the proof of Proposition 3.1.

Proposition C.1 (Alignment Sharpness of the Convex Surrogate)**.** *Suppose that $\mathcal{D}_{\mathbf{x}}$ is (L, R)-well-behaved, $\mathcal{U}_{(a,b)}$ is as in Definition 1.3, and $\epsilon, \delta > 0$. Let $\mu \gtrsim a^2LR^4/b$. Given any $\mathbf{w}^t \in \mathbb{B}(W)$, denote by $\hat{u}^t$ the optimal solution to (P) with respect to $\mathbf{w}^t$ and the sample set $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ drawn i.i.d. from $\mathcal{D}$. If m satisfies*

$$m \gtrsim dW^{9/2}b^4L^{-4}\log^4(d/(\epsilon\delta))(1/\epsilon^{3/2} + 1/(\epsilon\delta))\ ,$$

then, with probability at least $1-\delta$,

$$\begin{aligned}\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\cdot(\mathbf{w}^t - \mathbf{w}^*) \geq \mu\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2 \\ - 2(\mathrm{OPT} + \epsilon)/b - 2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2\ .\end{aligned}$$

Proof of Proposition 3.1. Given any weight parameter $\mathbf{w}^t \in \mathbb{B}(W)$ and $\hat{u}^t$ from the lemma statement, let u^t be the population-optimal activation, which recall was defined as the solution of the optimization problem (EP). Given a sample set $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$, suppose, hypothetically, that we could construct a new sample set S^* using $\mathbf{x}^{(i)}$'s from S but with the true labels without noise: $S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}_{i=1}^m$, $y^{*(i)} = u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$. For this reason, we define the following sub-problem, to use as reference:

$$\hat{u}^*_{\mathbf{w}} \in \operatorname*{argmin}_{u\in\mathcal{U}_{(a,b)}} \frac{1}{m}\sum_{i=1}^m (u(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{*(i)})^2. \tag{P*}$$

For a parameter $\mathbf{w}^t$, we will denote $\hat{u}^*_{\mathbf{w}^t}$ by $\hat{u}^{*t}$ for simplicity. Recall that the population version of $\hat{u}^{*t}$ was defined by (EP*).

To prove Proposition 3.1, we decompose $\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\cdot(\mathbf{w}^t - \mathbf{w}^*)$ into three summation terms, as follows.

$$\begin{aligned}&\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\cdot(\mathbf{w}^t - \mathbf{w}^*) \\ &= \frac{1}{m}\sum_{i=1}^m (\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - y^{(i)})(\mathbf{w}^t - \mathbf{w}^*)\cdot\mathbf{x}^{(i)} \\ &= \frac{1}{m}\sum_{i=1}^m \left((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - \hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t - \mathbf{w}^*)\cdot\mathbf{x}^{(i)} + (\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - y^{*(i)})(\mathbf{w}^t - \mathbf{w}^*)\cdot\mathbf{x}^{(i)}\right) \\ &\quad + \frac{1}{m}\sum_{i=1}^m (y^{*(i)} - y^{(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)} - \mathbf{w}^*\cdot\mathbf{x}^{(i)}).\end{aligned} \tag{11}$$

[2] A local error bound of this form was first used in (Zhang & Yin, 2013) under the name "restricted secant inequality."

We tackle each term in (11) separately, using the following arguments. Because the proofs of these claims are technical, we defer them to later subsections in Appendix C.

The first claim stated that the first summation in (11) is of order $(\sqrt{\epsilon}+\sqrt{\mathrm{OPT}})\|\mathbf{w}^t-\mathbf{w}^*\|_2+(\mathrm{OPT}+\epsilon)/b$ with high probability.

Claim C.2. Let $S=\{(\mathbf{x}^{(i)},y^{(i)})\}_{i=1}^m$ be m i.i.d. samples from $\mathcal{D}$ where m is as specified in the statement of Proposition 3.1. Let $\hat{u}^t$ be the solution of optimization problem (P) given $\mathbf{w}^t\in\mathbb{B}(W)$ and S. Furthermore, denote the uncorrupted version of S by $S^*=\{(\mathbf{x}^{(i)},y^{*(i)})\}_{i=1}^m$, where $y^{*(i)}=u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$. Let $\hat{u}^{*t}$ be the solution of problem (P*). Then, with probability at least $1-\delta$, it holds

$$\frac{1}{m}\sum_{i=1}^m((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\geq-(\sqrt{\epsilon}+\sqrt{\mathrm{OPT}})\|\mathbf{w}^t-\mathbf{w}^*\|_2-(\epsilon+\mathrm{OPT})/b\,.$$

The proof of Claim C.2 proceeds in the following routine: first, as we showed in Appendix F that due to some standard uniform convergence results the sample-optimal activations $\hat{u}^t$ and $\hat{u}^{*t}$ are close to their population counterparts, u^t and u^{*t}, in L_2^2 distance. Therefore, applying Chebyshev's inequality we are able to swap the sample-optimal activations in (11) by their population-optimal counterparts with high probability. Therefore, the proof boils down to lower bounding the following right-hand side:

$$\frac{1}{m}\sum_{i=1}^m((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\geq\mathrm{error}_1+\frac{1}{m}\sum_{i=1}^m(u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}.$$

Recall that in Lemma 3.3 we have shown that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]\leq\mathrm{OPT}$, thus using Chebyshev's inequality we can further show that the second term in the right-hand side of the inequality above can be lower bounded by some small error terms as well, completing the proof of the claim.

The second claim implemented the misalignment lemma (Lemma 3.2) and showed that the second summation term in (11) is basically some constant multiple of $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$, which is the main positive dominant factor among the three summations.

Claim C.3. Let $S^*=\{(\mathbf{x}^{(i)},y^{*(i)})\}_{i=1}^m$ be a sample set such that $\mathbf{x}^{(i)}$'s are m i.i.d. samples from $\mathcal{D}_\mathbf{x}$, and $y^{*(i)}=u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$ for each i. Let m be the value specified in the statement of Proposition 3.1. Then, given a parameter $\mathbf{w}^t\in\mathbb{B}(W)$, with probability at least $1-\delta$ it holds that

$$\frac{1}{m}\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\geq\frac{Ca^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2-\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2-\epsilon/b\,,$$

where C is an absolute constant.

The proof of Claim C.3 is rather technical. We first define an 'empirical inverse' of the activation u^*, and denote it by $\hat{f}$. Note that $u^*(z)\in\mathcal{U}_{(a,b)}$ is not necessarily strictly increasing when $z\leq0$, therefore $(u^*)^{-1}$ is not defined everywhere on $\mathbb{R}$, and the introduction of this 'empirical inverse' function $\hat{f}$ is needed. Then, adding and subtracting $\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))$ in the $\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\mathbf{w}^*\cdot\mathbf{x}^{(i)}$ term, we get

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\\
=&\frac{1}{m}\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})))\\
&+\frac{1}{m}\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)}).
\end{aligned}$$

Studying the KKT condition of the optimization problem (P*), we obtain a critical observation that the first term in the equation above is always positive. Then, observe that by our definition of $\hat{f}$ it works similar to the inverse of $(u^*)^{-1}$ and has the property that $|\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)}|\geq(1/b)|\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})|$ as well as

$(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}))(\hat{f}(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)})) - \mathbf{w}^* \cdot \mathbf{x}^{(i)}) \geq 0$, thus, the second term in the equation above can be lower bounded by $\frac{1}{bm}\sum_{i=1}^m (\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}))^2$. By some standard concentration techniques, the quantity above concentrates around its expectation $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]$, hence deploying the main structural result on alignment sharpness Proposition 3.1 completes the proof of Claim C.3.

Similar to Claim C.2, the last claim showed that the third summation term in (11) is of order $\sqrt{\mathrm{OPT}}\|\mathbf{w}^* - \mathbf{w}^t\|_2$, which is relative small compared to the positive term in Claim C.3.

Claim C.4. Let $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ be m i.i.d. samples from $\mathcal{D}$, and denote $S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}_{i=1}^m$ the uncorrupted version of S where $y^{*(i)} = u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$. Under the condition of Proposition 3.1, given a parameter $\mathbf{w}^t \in \mathbb{B}(W)$ with probability at least $1-\delta$ it holds:

$$\frac{1}{m}\sum_{i=1}^m (y^{*(i)} - y^{(i)})(\mathbf{w}^t \cdot \mathbf{x}^{(i)} - \mathbf{w}^* \cdot \mathbf{x}^{(i)}) \geq -\sqrt{\mathrm{OPT}}\|\mathbf{w}^* - \mathbf{w}^t\|_2 - (\mathrm{OPT} + \epsilon)/b\,.$$

The proof of Claim C.4 follows from a routine similar to Claim C.2.

Plugging the bounds from Claim C.2, Claim C.3, and Claim C.4 back into (11), we get that with probability at least $1-3\delta$,

$$\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) \geq \frac{Ca^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2 - 2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2 - 2(\mathrm{OPT} + \epsilon)/b,$$

for some absolute constant C, and the proof is now complete. □

C.2. Proof of Lemma 3.2

We restate and prove Lemma 3.2.

Lemma C.5 (Lower-Bounding L_2^2 Error by Misalignment)**.** *Let $u^* \in \mathcal{U}_{(a,b)}$, $\mathcal{D}_\mathbf{x}$ be (L,R)-well-behaved, and $f : \mathbb{R} \mapsto \mathbb{R}$ be square-integrable with respect to the measure of the distribution $\mathcal{D}_\mathbf{x}$. Then, for any $\mathbf{w}, \mathbf{w}^* \in \mathbb{R}^d$, it holds that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \gtrsim a^2LR^4\|(\mathbf{w}^*)^{\perp_\mathbf{w}}\|_2^2\,.$$

Proof. The statement holds trivially if $\mathbf{w}$ is parallel to $\mathbf{w}^*$, so assume this is not the case. Let $\mathbf{v} = (\mathbf{w}^*)^{\perp_\mathbf{w}} = \mathbf{w}^* - (\mathbf{w}^* \cdot \mathbf{w})\mathbf{w}/\|\mathbf{w}\|_2^2$. Suppose first that $\mathbf{w}\cdot\mathbf{w}^* \geq 0$. Then $\mathbf{w}^* = \alpha\mathbf{w} + \mathbf{v}$, for some $\alpha > 0$. Let V be the subspace spanned by $\mathbf{w}, \mathbf{v}$. Then,

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{w}\cdot\mathbf{x}_V) - u^*(\mathbf{w}^*\cdot\mathbf{x}_V))^2] \geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{w}\cdot\mathbf{x}_V) - u^*(\mathbf{w}^*\cdot\mathbf{x}_V))^2\mathbb{1}\{\mathbf{x}_V \in A\}]\,,$$

for any $A \subseteq \mathbb{R}^d$. For ease of notation, we drop the subscript V, and we assume that all $\mathbf{x}$ are projected to the subspace V. We denote by $\tilde{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|_2$ (resp. $\tilde{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_2$) the unit vector in the direction of $\mathbf{w}$ (resp. $\mathbf{v}$). We choose $A = \{\mathbf{w}\cdot\mathbf{x} \geq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (R/16, R/8) \cup (3R/8, R/2)\}$.

The idea of the proof is to utilize the non-decreasing property of u^* and the fact that the marginal distribution $\mathcal{D}_\mathbf{x}$ is anti-concentrated on the subspace S. In short, for any $|\tilde{\mathbf{v}}\cdot\mathbf{x}| \leq R$, by the non-decreasing property of u^* we know that $f(\mathbf{w}\cdot\mathbf{x})$ falls into one of the following four intervals:

$$(-\infty,\ u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2R/32)], \qquad (u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2R/32),\ u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2R/4)],$$
$$(u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2R/4),\ u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2R)], \qquad (u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2R),\ +\infty).$$

When $f(\mathbf{w}\cdot\mathbf{x})$ belongs to any of the intervals above, we can show that with some constant probability, the difference between $\mathbf{w}^*\cdot\mathbf{x}$ and $\mathbf{w}\cdot\mathbf{x}$ is proportional to $\|\mathbf{v}\|_2$ and hence $u^*(\mathbf{w}^*\cdot\mathbf{x})$ is far away from $f(\mathbf{w}\cdot\mathbf{x})$, due to the well-behaved property of the marginal $\mathcal{D}_\mathbf{x}$.

To indicate that $f(\mathbf{w}\cdot\mathbf{x})$ belongs to one of the intervals above, denote

$$I_1(\mathbf{x}) = f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2R/32),$$
$$I_2(\mathbf{x}) = f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2R/4),$$
$$I_3(\mathbf{x}) = f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2R).$$

For any $\mathbf{x} \in \mathbb{R}^d$, using the assumption that u^* is non-decreasing, we have that $I_1(\mathbf{x}) \geq I_2(\mathbf{x}) \geq I_3(\mathbf{x})$; as a consequence, it must be that $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$ or $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$.

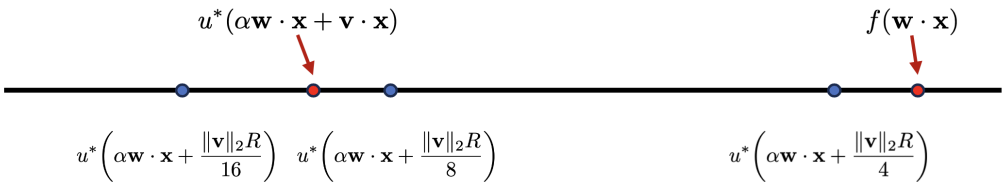

Figure 1: *Under the assumption that* $\tilde{\mathbf{v}}\cdot\mathbf{x} \in (R/16, R/8)$*, and* $I_1(\mathbf{x}) \geq 0, I_2(\mathbf{x}) \geq 0$*, the distance between* $f(\mathbf{w}\cdot\mathbf{x})$ *and* $u^*(\mathbf{w}^*\cdot\mathbf{x})$ *is at least* $|u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^*\cdot\mathbf{x})| \geq a\|\mathbf{v}\|_2 R/8$.

Case 1: $f(\mathbf{w}\cdot\mathbf{x}) \in (u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/4),\, \infty)$. Then $I_1(\mathbf{x}) \geq I_2(\mathbf{x}) \geq 0$. Let $B := \{\mathbf{w}\cdot\mathbf{x} \geq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (R/16, R/8)\}$ and notice that $B \subseteq A$. We have that when $\mathbf{x} \in B$,

$$u^*(\mathbf{w}^*\cdot\mathbf{x}) = u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2\tilde{\mathbf{v}}\cdot\mathbf{x}) \in (u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/16),\; u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/8)),$$

thus we can conclude that

$$\begin{aligned}
&(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\} \\
&= \left(\{f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/4)\} + \{u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^*\cdot\mathbf{x})\}\right)^2\mathbb{1}\{\mathbf{x}\in B\} \\
&\geq (u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\}\,,
\end{aligned}$$

where in the last inequality we used that $I_2(\mathbf{x}) = f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/4) \geq 0$ and $u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^*\cdot\mathbf{x}) \geq 0$ by the non-decreasing property of u^* and that if $a, b \geq 0$ then $(a+b)^2 \geq \max(a,b)^2$. Further, using $u^*(t) - u^*(t') \geq a(t-t')$ for $t \geq t' \geq 0$ (which holds by assumption) and $\mathbf{w}^* = \alpha\mathbf{w} + \mathbf{v}$, we have

$$\begin{aligned}
(u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\} &\geq a^2(\|\mathbf{v}\|_2 R/4 - \mathbf{v}\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{x}\in B\} \\
&\geq a^2\|\mathbf{v}\|_2^2(R/8)^2\mathbb{1}\{\mathbf{x}\in B\}\,,
\end{aligned}$$

where in the last inequality we used that $0 \leq \tilde{\mathbf{v}}\cdot\mathbf{x} \leq R/8$ (by the definition of the event B). A visual explanation of the result above is displayed in Figure 1.

Case 2: $f(\mathbf{w}\cdot\mathbf{x}) \in (-\infty,\, u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/32))$. Then $0 \geq I_1(\mathbf{x}) \geq I_2(\mathbf{x})$. We follow a similar argument as in the previous case. In particular, we begin with

$$\begin{aligned}
&(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\} \\
&= \left(\{f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/32)\} + \{u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/32) - u^*(\mathbf{w}^*\cdot\mathbf{x})\}\right)^2\mathbb{1}\{\mathbf{x}\in B\}. \qquad (12)
\end{aligned}$$

Note that $I_1(\mathbf{x}) \leq 0$ and $u^*(\mathbf{w}^*\cdot\mathbf{x}) = u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2\tilde{\mathbf{v}}\cdot\mathbf{x}) \geq u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/32)$ since $\tilde{\mathbf{v}}\cdot\mathbf{x} \geq R/16 \geq R/32$ for $\mathbf{x} \in B$, thus the two terms in curly brackets in (12) have the same sign and we further have:

$$\begin{aligned}
(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\} &\geq (u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2 R/32) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\} \\
&\geq a^2\|\mathbf{v}\|_2^2(R/32)^2\mathbb{1}\{\mathbf{x}\in B\},
\end{aligned}$$

where in the first inequality we used the fact that $(a+b)^2 \geq \max\{a^2, b^2\}$ when both $a, b \leq 0$.

By Case 1 and Case 2, we can conclude that when $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$, it must be:

$$(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\} \geq a^2\|\mathbf{v}\|_2^2 R^2/2^{10}\mathbb{1}\{\mathbf{x}\in B\}. \qquad (13)$$

Case 3: $f(\mathbf{w}\cdot\mathbf{x}) \in (u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R), +\infty)$. Then $I_2(\mathbf{x}) \geq I_3(\mathbf{x}) \geq 0$ and we choose $B' = \{\mathbf{w}\cdot\mathbf{x} \geq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (3R/8, R/2)\}$. Following the same reasoning as in the previous two cases, we have

$$\begin{aligned}
&(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&= \big(\{f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R)\} + \{u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R) - u^*(\mathbf{w}^*\cdot\mathbf{x})\}\big)^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&\geq (u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&\geq a^2\|\mathbf{v}\|_2^2 (R/2)^2 \mathbb{1}\{\mathbf{x}\in B'\}.
\end{aligned}$$

Case 4: $f(\mathbf{w}\cdot\mathbf{x}) \in (-\infty, u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/4))$. Then $0 \geq I_2(\mathbf{x}) \geq I_3(\mathbf{x})$. It follows that

$$\begin{aligned}
&(f(\mathbf{w}\cdot\mathbf{x}) - u(\mathbf{w}^*\cdot\mathbf{x}))^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&= \big(\{f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/4)\} + \{u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/4) - u(\mathbf{w}^*\cdot\mathbf{x})\}\big)^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&\geq a^2\|\mathbf{v}\|_2^2 (R/8)^2 \mathbb{1}\{\mathbf{x}\in B'\}.
\end{aligned}$$

Thus, we conclude from Case 3 and Case 4 that when $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$, we have

$$(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2 \mathbb{1}\{\mathbf{x}\in B'\} \geq a^2\|\mathbf{v}\|_2^2 (R^2/64) \mathbb{1}\{\mathbf{x}\in B'\}. \tag{14}$$

Recall that for any $\mathbf{x}$, at least one of the inequalities $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$ or $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$ happens, thus, $\mathbb{1}\{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\} \geq 1 - \mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}$. Therefore, the probability mass of the region $(B \cap \{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\}) \cup (B' \cap \{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\})$ can be lower bounded by:

$$\begin{aligned}
&\mathbf{Pr}\Big[\mathbf{x} \in (B \cap \{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\}) \cup (B' \cap \{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\})\Big] \\
&= \int_V \Big(\mathbb{1}\{\mathbf{x}\in B\}\mathbb{1}\{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\} + \mathbb{1}\{\mathbf{x}\in B'\}\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\Big)\gamma(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&\geq \int_{V, \|\mathbf{x}\|_\infty \leq R} \Big(\mathbb{1}\{\mathbf{x}\in B\}\mathbb{1}\{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\} + \mathbb{1}\{\mathbf{x}\in B'\}\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\Big) L\,\mathrm{d}\mathbf{x} \\
&\geq L\int_{V, \|\mathbf{x}\|_\infty \leq R} \Big(\mathbb{1}\{\mathbf{x}\in B\} + (\mathbb{1}\{\mathbf{x}\in B'\} - \mathbb{1}\{\mathbf{x}\in B\})\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\Big)\,\mathrm{d}\mathbf{x},
\end{aligned} \tag{15}$$

where in the first inequality we used the assumption that $\mathcal{D}_\mathbf{x}$ is (L, R)-well-behaved. As a visual explanation of the lower bound above, we include the following Figure 2.

To finish bounding the probability in (15), it remains to bound the integral from its final inequality, which now does not involve the pdf anymore, as we used the anti-concentration property of $\mathcal{D}_\mathbf{x}$ to uniformly bound below $\gamma(\mathbf{x})$. Recall that by definition, $I_1(\mathbf{x}), I_2(\mathbf{x}), I_3(\mathbf{x})$ are functions of $\mathbf{w}\cdot\mathbf{x}$ that do not depend on $\tilde{\mathbf{v}}\cdot\mathbf{x}$. Denote the projection of $\mathbf{x}$ on the standard basis of space V by $\mathbf{x}_{\tilde{\mathbf{w}}} = \tilde{\mathbf{w}}\cdot\mathbf{x}$ and $\mathbf{x}_{\tilde{\mathbf{v}}} = \tilde{\mathbf{v}}\cdot\mathbf{x}$. Then, we have:

$$\begin{aligned}
&\int_{V, \|\mathbf{x}\|_\infty \leq R} \Big(\mathbb{1}\{\mathbf{x}\in B'\} - \mathbb{1}\{\mathbf{x}\in B\}\Big)\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\,\mathrm{d}\mathbf{x} \\
&= \int_{|\mathbf{x}_{\tilde{\mathbf{w}}}| \leq R}\int_{|\mathbf{x}_{\tilde{\mathbf{v}}}| \leq R} \left(\mathbb{1}\left\{\mathbf{x}_{\tilde{\mathbf{v}}} \in \left(\frac{3R}{8}, \frac{R}{2}\right)\right\} - \mathbb{1}\left\{\mathbf{x}_{\tilde{\mathbf{v}}} \in \left(\frac{R}{16}, \frac{R}{8}\right)\right\}\right)\mathrm{d}\mathbf{x}_{\tilde{\mathbf{v}}}\,\mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{w}}} \geq 0, I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\,\mathrm{d}\mathbf{x}_{\tilde{\mathbf{w}}} \\
&= \int_{|\mathbf{x}_{\tilde{\mathbf{w}}}| \leq R} \mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{w}}} \geq 0, I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\,\mathrm{d}\mathbf{x}_{\tilde{\mathbf{w}}} \int_{|\mathbf{x}_{\tilde{\mathbf{v}}}| \leq R} \left(\mathbb{1}\left\{\mathbf{x}_{\tilde{\mathbf{v}}} \in \left(\frac{3R}{8}, \frac{R}{2}\right)\right\} - \mathbb{1}\left\{\mathbf{x}_{\tilde{\mathbf{v}}} \in \left(\frac{R}{16}, \frac{R}{8}\right)\right\}\right)\mathrm{d}\mathbf{x}_{\tilde{\mathbf{v}}} \\
&\geq 0,
\end{aligned}$$

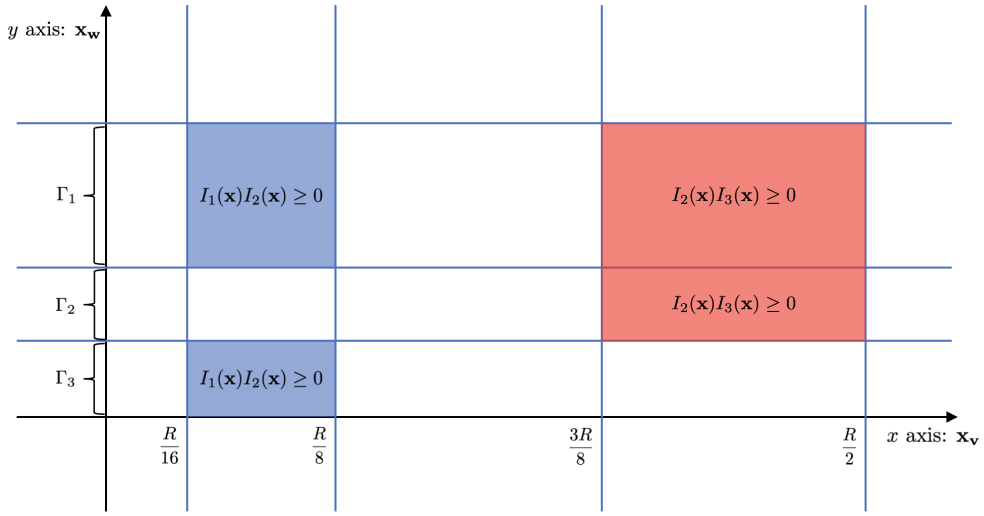

Figure 2: *On the 2-dimensional space V spanned by $(\mathbf{x}_\mathbf{v}, \mathbf{x}_\mathbf{w})$, at each point $\mathbf{x} \in B \cup B'$, it must be that $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$ or $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$. Γ_1 denotes the interval of $\mathbf{x}_\mathbf{w} = \mathbf{w} \cdot \mathbf{x}$ such that $f(\mathbf{w} \cdot \mathbf{x}) \geq u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R)$, hence both $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$, $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$; Γ_2 denotes the interval of $\mathbf{x}_\mathbf{w}$ such that $f(\mathbf{w} \cdot \mathbf{x}) \in (u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/32), u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4))$, hence $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$; finally, Γ_3 denotes the interval of $\mathbf{x}_\mathbf{w}$ such that $f(\mathbf{w} \cdot \mathbf{x}) \in (u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4), u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/))$, hence $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$. The area of the union of the red and blue regions is the lower bound on the probability in (15). As displayed in the figure, the sum of the blue and red region is lower bounded by $\mathbb{1}\{\mathbf{x} \in B\} + (\mathbb{1}\{\mathbf{x} \in B'\} - \mathbb{1}\{\mathbf{x} \in B\})\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}$.*

Plugging the inequality above back into (15), we get:

$$\begin{aligned}
&\mathbf{Pr}\left[\mathbf{x} \in B \cap \{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\} + B' \cap \{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\right] \\
&\geq L \int_{V, \|\mathbf{x}\|_\infty \leq R} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \geq 0, \tilde{\mathbf{v}} \cdot \mathbf{x} \in (R/16, R/8)\} \, \mathrm{d}\mathbf{x} \\
&= L \iint (\mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{w}}} \in (0, R)\} \, \mathrm{d}\mathbf{x}_{\tilde{\mathbf{w}}}) \mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{v}}} \in (R/16, R/8)\} \, \mathrm{d}\mathbf{x}_{\tilde{\mathbf{v}}} = LR^2/16. \qquad (16)
\end{aligned}$$

We are now ready to provide a lower bound on the L_2^2 distance between $f(\mathbf{w} \cdot \mathbf{x})$ and $u^*(\mathbf{w}^* \cdot \mathbf{x})$. Combining the inequalities from (13) and (14), we get

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(f(\mathbf{w} \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \\
&\geq \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(f(\mathbf{w} \cdot \mathbf{x}_V) - u^*(\mathbf{w}^* \cdot \mathbf{x}_V))^2 \mathbb{1}\{\mathbf{x}_V \in A\}] \\
&\geq a^2 (R^2/1024) \|\mathbf{v}\|_2^2 \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [\mathbb{1}\{\{\mathbf{x}_V \in B \cap \{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\}\} \cup \{B' \cap \{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\}\}] \\
&\geq a^2 (R^4/2^{13}) L \|\mathbf{v}\|_2^2 \,,
\end{aligned}$$

where we used (16) in the last inequality.

Now for the case where $\mathbf{w} \cdot \mathbf{w}^* \leq 0$, it holds $\mathbf{w}^* = \alpha \mathbf{w} + \mathbf{v}$ with $\alpha \leq 0$. Considering instead $A = \{\mathbf{w} \cdot \mathbf{x} \leq 0, \tilde{\mathbf{v}} \cdot \mathbf{x} \in (R/16, R/8) \cup (3R/8, R/2)\}$ and similarly $B = \{\mathbf{w} \cdot \mathbf{x} \leq 0, \tilde{\mathbf{v}} \cdot \mathbf{x} \in (R/16, R/8)\}$, $B' = \{\mathbf{w} \cdot \mathbf{x} \leq 0, \tilde{\mathbf{v}} \cdot \mathbf{x} \in (R/3, R/2)\}$, then all the steps above remains valid without modification. This completes the proof of Lemma 3.2. □

C.3. Proof of Lemma 3.3

In this section, we restate and prove Lemma 3.3. We first show the following claim, which is inspired by Lemma 9 in (Kakade et al., 2011).

Claim C.6. Let $\mathbf{w}^t \in \mathbb{B}(W)$ and let u^{*t}, u^t be defined as solutions to (EP*), (EP), respectively. It holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - v(\mathbf{w}^t\cdot\mathbf{x}))(y - u^t(\mathbf{w}^t\cdot\mathbf{x}))] \geq 0$, for any $v \in \mathcal{U}_{(a,b)}$; similarly, it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - v'(\mathbf{w}^t\cdot\mathbf{x}))(y^* - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))] \geq 0$, for any $v' \in \mathcal{U}_{(a,b)}$.

Proof of Claim C.6. The proof is Let us denote by $\mathcal{F}_t$ the set of functions of the form $f(\mathbf{x}) = u(\mathbf{w}^t\cdot\mathbf{x})$, where $u \in \mathcal{U}_{(a,b)}$ and $\mathbf{w}^t$ is a fixed vector in $\mathbb{B}(W)$. We first observe that $\mathcal{F}_t$ is a convex set of functions. This is because for any $\alpha \in [0,1]$, for any $f_1, f_2 \in \mathcal{F}_t$ such that $f_1(\mathbf{x}) = u_1(\mathbf{w}^t\cdot\mathbf{x})$, $f_2 = u_2(\mathbf{w}^t\cdot\mathbf{x})$, let $u_3(\cdot) = \alpha u_1(\cdot) + (1-\alpha)u_2(\cdot)$, it holds:

$$\alpha f_1(\mathbf{x}) + (1-\alpha)f_2(\mathbf{x}) = \alpha u_1(\mathbf{w}^t\cdot\mathbf{x}) + (1-\alpha)u_2(\mathbf{w}^t\cdot\mathbf{x}) = u_3(\mathbf{w}^t\cdot\mathbf{x}),$$

note that u_3 is also (a,b)-bounded, non-decreasing, and $u_3(0) = 0$, hence $u_3 \in \mathcal{U}_{(a,b)}$ and $f_3(\mathbf{x}) = u_3(\mathbf{w}^t\cdot\mathbf{x}) \in \mathcal{F}_t$, thus, $\mathcal{F}_t$ is convex.

Since $\mathcal{F}_t$ is a convex set of functions, essentially we can regard $u^t(\mathbf{w}^t\cdot\mathbf{x})$ as the L_2 projection of y (which is a function of $\mathbf{x}$) onto the convex set $\mathcal{F}_t$. Classic inequalities of ℓ_2 projection can be seamlessly transformed in our case. In particular, below we prove that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - v(\mathbf{w}^t\cdot\mathbf{x}))(y - u^t(\mathbf{w}^t\cdot\mathbf{x}))] \geq 0, \forall v \in \mathcal{U}_{(a,b)}. \tag{17}$$

To prove (17), note first that $f_u(\mathbf{x}) = u^t(\mathbf{w}^t\cdot\mathbf{x}) \in \mathcal{F}_t$ and $f_v(\mathbf{x}) = v(\mathbf{w}^t\cdot\mathbf{x}) \in \mathcal{F}_t$ since $u^t, v \in \mathcal{U}_{(a,b)}$. Thus, for any $\alpha \in (0,1)$, we have $\alpha f_v(\mathbf{x}) + (1-\alpha)f_u(\mathbf{x}) \in \mathcal{F}_t$. Furthermore, by definition of u^t, $\forall f \in \mathcal{F}_t$ we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - y)^2] \leq \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - y)^2]$, therefore, it holds:

$$\begin{aligned}
0 &\leq \frac{1}{\alpha}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\alpha f_v(\mathbf{x}) + (1-\alpha)f_u(\mathbf{x}) - y)^2] - \frac{1}{\alpha}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - y)^2] \\
&= \frac{1}{\alpha}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - y + \alpha(v(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x})))^2 - (u^t(\mathbf{w}^t\cdot\mathbf{x}) - y)^2] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[2(u^t(\mathbf{w}^t\cdot\mathbf{x}) - y)(v(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x})) + \alpha(v(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2].
\end{aligned}$$

Let $\alpha \downarrow 0$, and note that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(v(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2] < +\infty$, we thus have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - v(\mathbf{w}^t\cdot\mathbf{x}))(y - u^t(\mathbf{w}^t\cdot\mathbf{x}))] \geq 0,$$

proving the claim.

We can show that a similar result also holds for u^{*t} and y^*. Specifically, we have:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - v'(\mathbf{w}^t\cdot\mathbf{x}))(y^* - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))] \geq 0, \forall v' \in \mathcal{U}_{(a,b)}. \tag{18}$$

This completes the proof of Claim C.6. □

We now proceed to the proof of Lemma 3.3.

Lemma C.7 (Closeness of Population-Optimal Activations)**.** *Let $\mathbf{w}^t \in \mathbb{B}(W)$ and let u^{*t}, u^t be defined as solutions to (EP*), (EP), respectively. Then,*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] \leq \mathrm{OPT}.$$

Proof. Summing up the first and second statement of Claim C.6 (i.e., (17) and (18)) with $v = u^{*t} \in \mathcal{U}_{(a,b)}$ in (17) and $v' = u^t \in \mathcal{U}_{(a,b)}$ in (18), we get:

$$\begin{aligned}
0 &\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(y - u^t(\mathbf{w}^t\cdot\mathbf{x})) + (u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))(y^* - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(y - y^* + u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(y - y^*)] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]
\end{aligned}$$

Therefore, moving the second term above to the left-hand side, then applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] &\le \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(y-y^*)]\\ &\le \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]\,\mathbf{E}[(y-y^*)^2]},\end{aligned}$$

completing the proof of Lemma 3.3. □

C.4. Proof of Corollary 3.4

Corollary C.8 (Closeness of Idealized and Attainable Activations)**.** *Let $\epsilon,\delta>0$. Given a parameter $\mathbf{w}^t\in\mathbb{B}(W)$ and $m\gtrsim d\log^4(d/(\epsilon\delta))(b^2W^3/(L^2\epsilon))^{3/2}$ samples from $\mathcal{D}$, let $\hat{u}^t$ be the sample-optimal activation on these samples given $\mathbf{w}^t$, as defined in (P). Then, with probability at least $1-\delta$,*

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\\ &\qquad\le 3(\epsilon+\mathrm{OPT}+b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2).\end{aligned}$$

Proof. The corollary follows directly from the combination of Lemma F.4 and Lemma 3.3, as we have:

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\\ &=\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x})+u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x})+u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\\ &\le 3(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2]+\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]+\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2])\\ &\le 3(\epsilon+\mathrm{OPT}+b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2),\end{aligned}$$

where we used that because $u^{*t}\in\mathrm{argmin}_{u\in\mathcal{U}_{(a,b)}}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]$, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\le\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^*(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\le b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2$, with the last inequality following from the fact that $u^*\in\mathcal{U}_{(a,b)}$. □

C.5. Proof of Claim C.2

In this subsection, we prove Claim C.2 that appeared in Appendix C.1, the proof of Proposition 3.1.

Claim C.9. Let $S=\{(\mathbf{x}^{(i)},y^{(i)})\}_{i=1}^m$ be m i.i.d. samples from $\mathcal{D}$ where m is as specified in the statement of Proposition 3.1. Let $\hat{u}^t$ be the solution of optimization problem (P) given $\mathbf{w}^t\in\mathbb{B}(W)$ and S. Furthermore, denote the uncorrupted version of S by $S^*=\{(\mathbf{x}^{(i)},y^{*(i)})\}_{i=1}^m$, where $y^{*(i)}=u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$. Let $\hat{u}^{*t}$ be the solution of problem (P*). Then, with probability at least $1-\delta$, it holds

$$\frac{1}{m}\sum_{i=1}^m((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\ge-(\sqrt{\epsilon}+\sqrt{\mathrm{OPT}})\|\mathbf{w}^t-\mathbf{w}^*\|_2-(\epsilon+\mathrm{OPT})/b\,.$$

Proof. Adding and subtracting $u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})$ and $u^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})$, we have

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\\ &=\frac{1}{m}\sum_{i=1}^m(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}+\frac{1}{m}\sum_{i=1}^m(u^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\\ &\quad+\frac{1}{m}\sum_{i=1}^m(u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}.\end{aligned}\tag{19}$$

To proceed, we use that both $\hat{u}^{*t}(z)$ and $\hat{u}^t(z)$ are close to their population counterparts $u^t(z)$ and $u^{*t}(z)$, respectively. In particular, in Lemma F.4 and Lemma F.2, we showed that using a dataset S of m samples such that

$$m\gtrsim d\log^4(d/(\epsilon\delta))\left(\frac{b^2W^3}{L^2\epsilon}\right)^{3/2},$$

we have that with probability at least $1-\delta$, for all $\mathbf{w}^t, \mathbf{w}^* \in \mathbb{B}(W)$ it holds

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2] \le \epsilon, \ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] \le \epsilon. \tag{20}$$

Now suppose that the inequalities in (20) hold for the given $\mathbf{w}^t \in \mathbb{B}(W)$ (which happens with probability at least $1-\delta$). Applying Chebyshev's inequality to the first summation term in (19), we get:

$$\begin{aligned}\mathbf{Pr}\left[\left|\frac{1}{m}\sum_{i=1}^m(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}]\right| \ge s\right] \\ \le \frac{1}{ms^2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2],\end{aligned} \tag{21}$$

since $\mathbf{x}^{(i)}$ are i.i.d random variables. The next step is to bound the variance. Note that $\mathcal{D}_\mathbf{x}$ possesses a $1/L$-sub-exponential tail, thus we have $\mathbf{Pr}[|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}| \ge \|\mathbf{w}^t-\mathbf{w}^*\|_2 r] \le (2/L^2)\exp(-Lr)$. Choose $r = \frac{2W}{L}\log(2/(L^2\epsilon'))$; then, we have $\mathbf{Pr}[|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}| \ge r] \le \epsilon'$. Now we separate the variance under two events: $A = \{\mathbf{x} : |(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}| \le r\}$ and $\neg A = \{\mathbf{x} : |(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}| \ge r\}$.

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{A\}] \\ &\quad + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2(1-\mathbb{1}\{A\})].\end{aligned} \tag{22}$$

Using that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2] \le \epsilon$, the first term in (22) can be bounded as follows:

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{A\}] &\le r^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2] \\ &\le r^2\epsilon = \frac{4W^2\epsilon}{L^2}\log^2(2/(L^2\epsilon')).\end{aligned} \tag{23}$$

The second term in (22) can be bounded using that both $\hat{u}^t$ and u^t are non-decreasing b-Lipschitz and vanish at zero (thus $|\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})| \le b|\mathbf{w}^t\cdot\mathbf{x}|$ and $|u^t(\mathbf{w}^t\cdot\mathbf{x})| \le b|\mathbf{w}^t\cdot\mathbf{x}|$, with their signs determined by the sign of $\mathbf{w}^t\cdot\mathbf{x}$) and then applying Young's inequality:

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2(1-\mathbb{1}\{A\})] \\ &\le b^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}^t\cdot\mathbf{x})^2(\mathbf{w}^t\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2(1-\mathbb{1}\{A\})] \\ &\le 2b^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t\cdot\mathbf{x})^4 + (\mathbf{w}^t\cdot\mathbf{x})^2(\mathbf{w}^*\cdot\mathbf{x})^2)(1-\mathbb{1}\{A\})].\end{aligned}$$

Since $\mathcal{D}_\mathbf{x}$ is sub-exponential, we have $\mathbf{E}[(\mathbf{v}\cdot\mathbf{x})^8] \le c^2/L^8$ for some absolute constant c, hence

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}^t\cdot\mathbf{x})^4(1-\mathbb{1}\{A\})] \le \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[W^8((\mathbf{w}^t/\|\mathbf{w}^t\|_2)\cdot\mathbf{x})^8]\,\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}| \ge r]} \le cW^4\sqrt{\epsilon'}/L^4.$$

Similarly, for $\mathbf{E}[(\mathbf{w}^t\cdot\mathbf{x})^2(\mathbf{w}^*\cdot\mathbf{x})^2(1-\mathbb{1}\{A\})]$, we have:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}^t\cdot\mathbf{x})^2(\mathbf{w}^*\cdot\mathbf{x})^2(1-\mathbb{1}\{A\})] \le 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t\cdot\mathbf{x})^4 + (\mathbf{w}^*\cdot\mathbf{x})^4)(1-\mathbb{1}\{A\})] \le 2c(W/L)^4\sqrt{\epsilon'}.$$

Combining the inequalities above with (23), we get the final upper bound on the variance in (22):

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2] \le \frac{4W^2\epsilon}{L^2}\log^2(2/(L^2\epsilon')) + 6cb^2(W/L)^4\sqrt{\epsilon'}.$$

Thus, choosing $s = \epsilon/b$ in (21), $\epsilon' = \epsilon^2$, and using $m \gtrsim W^4b^4\log^2(1/\epsilon)/(\epsilon\delta L^4)$ samples we get

$$\frac{1}{ms^2}\left(\frac{4W^2\epsilon}{L^2}\log^2(2/(L\epsilon')) + \frac{12cb^2W^4\sqrt{\epsilon'}}{L^4}\right) \lesssim \frac{b^2L^4\epsilon\delta}{\epsilon^2W^4b^4\log^2(1/\epsilon)}\left(\frac{W^2\epsilon}{L^2}\log^2\left(\frac{1}{L\epsilon}\right) + \frac{b^2W^4\epsilon}{L^4}\right) \le \delta\,.$$

Plugging the inequality above back into (21) and recalling that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2]\le\epsilon$ (from (20)), we finally have with probability at least $1-\delta$,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\\
&\ge \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}]-\epsilon/b\\
&\ge -\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2]}-\epsilon/b\\
&\ge -\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2-\epsilon/b,
\end{aligned}$$

where in the second inequality we used the Cauchy-Schwarz inequality and in the last inequality we used the assumption that $\mathcal{D}_\mathbf{x}$ is isotropic, i.e., $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x}\mathbf{x}^\top]=\mathbf{I}$. Finally, note that (20) holds with probability at least $1-\delta$, applying a union bound we get that with probability at least $1-2\delta$, we have

$$\frac{1}{m}\sum_{i=1}^m(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\ge-\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2-\epsilon/b.$$

In summary, to guarantee that the inequality above remains valid, we need the batch size to be:

$$m\gtrsim\frac{dW^{9/2}b^4\log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}}+\frac{1}{\epsilon\delta}\right).\tag{24}$$

We finished bounding the first term in (19).

Since the same statements hold for the relationship between $\hat{u}^{*t}$ and u^{*t} as they do for $\hat{u}^t$ and u^t, using the same argument we also get that with probability at least $1-2\delta$,

$$\frac{1}{m}\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\ge-\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2-\epsilon/b,$$

which is the lower bound for the second term in (19).

Lastly, for the third term in (19), since in Lemma 3.3 we showed that for any $\mathbf{w}^t$ it always holds:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]\le\mathrm{OPT},$$

the only change of the previous steps is at the right-hand side of (23), where instead of having the upper bound of $r^2\epsilon$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{A\}]\le r^2\mathrm{OPT}=\frac{4W^2\mathrm{OPT}}{L^2}\log^2(2/(L^2\epsilon')).$$

And in the same reason, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2(1-\mathbb{1}\{A\})]\le 6cb^2(W/L)^4\sqrt{\epsilon'}.$$

As a result, Chebyshev's inequality yields:

$$\begin{aligned}
&\mathbf{Pr}\left[\left|\frac{1}{m}\sum_{i=1}^m(u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}]\right|\ge s\right]\\
&\quad\le\frac{1}{ms^2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2]\\
&\quad\le\frac{1}{ms^2}\left(\frac{4W^2\mathrm{OPT}}{L^2}\log^2(2/(L^2\epsilon'))+\frac{6cb^2W^4\sqrt{\epsilon'}}{L^4}\right).
\end{aligned}$$

Now instead of choosing $s = \epsilon$, we let $s = (\mathrm{OPT} + \epsilon)/b$ and keep ϵ' as ϵ^2 to get

$$\frac{1}{ms^2}\left(\frac{4W^2\mathrm{OPT}}{L^2}\log^2\left(\frac{2}{L^2\epsilon'}\right) + \frac{12cb^2W\sqrt{\epsilon'}}{L^4}\right)$$
$$\lesssim \frac{b^2L^4\epsilon\delta}{dW^{9/2}b^4\log^4(d/(\epsilon\delta))(\mathrm{OPT}+\epsilon)^2}\left(\frac{W^2\mathrm{OPT}}{L^2}\log^2\left(\frac{1}{L\epsilon}\right) + \frac{b^2W^4\epsilon}{L^4}\right) \le \delta,$$

under our choice of m as specified in (24). Thus, we have that with probability at least $1-\delta$, it holds

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}(u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\\
&\ge \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}] - (\mathrm{OPT}+\epsilon)/b\\
&\ge -\sqrt{\mathrm{OPT}}\|\mathbf{w}^t-\mathbf{w}^*\|_2 - (\mathrm{OPT}+\epsilon)/b,
\end{aligned}$$

where in the last inequality we used the fact that

$$\begin{aligned}
|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}]| &\le \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x})^2]}\\
&\le \sqrt{\mathrm{OPT}}\|\mathbf{w}^t-\mathbf{w}^*\|_2,
\end{aligned}$$

since $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] \le \mathrm{OPT}$ by Lemma 3.3.

Therefore, combing the upper bounds on the three terms in (19) we get that with probability at least $1-5\delta$, it holds:

$$\frac{1}{m}\sum_{i=1}^{m}((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - \hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} \ge -(2\sqrt{\epsilon}+\sqrt{\mathrm{OPT}})\|\mathbf{w}^t-\mathbf{w}^*\|_2 - (3\epsilon+\mathrm{OPT})/b. \tag{25}$$

Since (25) was proved using arbitrary $\epsilon, \delta > 0$, it remains to replace $\delta \leftarrow \delta/5$ and $\epsilon \leftarrow \epsilon/4$ to complete the proof of Claim C.2. □

C.6. Proof of Claim C.3

In this subsection, we prove Claim C.3 that appeared in the proof of Proposition 3.1 in Appendix C.1.

Claim C.10. Let $S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}_{i=1}^m$ be a sample set such that $\mathbf{x}^{(i)}$'s are m i.i.d. samples from $\mathcal{D}_\mathbf{x}$, and $y^{*(i)} = u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$ for each i. Let m be the value specified in the statement of Proposition 3.1. Then, given a parameter $\mathbf{w}^t \in \mathbb{B}(W)$, with probability at least $1-\delta$ it holds that

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - y^{*(i)})(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} \ge \frac{Ca^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2 - \sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2 - \epsilon/b\,,$$

where C is an absolute constant.

Proof. Before we proceed to the proof of the claim, let us consider first the inverse of u^*. Since $u^*(z) \in \mathcal{U}_{(a,b)}$ is strictly increasing when $z \ge 0$, hence $(u^*)^{-1}(\alpha)$ exists for $\alpha \ge 0$. However, when $z \le 0$, $u^*(z)$ could be a constant on some intervals, hence $(u^*)^{-1}(\alpha)$ might not exist for every $\alpha \le 0$. We consider instead an 'empirical' version of $(u^*)^{-1}(\alpha)$ based on S^*, which is defined on every $\alpha \in \mathbb{R}$. Given a sample set $S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}$ where $y^{*(i)} = u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$, let us sort the the index i in the increasing order of $\mathbf{w}^*\cdot\mathbf{x}^{(i)}$, i.e., $\mathbf{w}^*\cdot\mathbf{x}^{(1)} \le \cdots \le \mathbf{w}^*\cdot\mathbf{x}^{(m)}$. Since u^* is a monotone function, this implies $y^{*(i)}$'s are also in increasing order, i.e., we have $y^{*(1)} \le \cdots \le y^{*(m)}$. We then partition the set $\{y^{*(i)}\}_{i=1}^m$ into blocks

$$\Delta_s = \{y^{*(k_{s-1}+1)}, \ldots, y^{*(k_s)}\},\ \text{s.t. } y^{*(k_{s-1}+1)} = \cdots = y^{*(k_s)} = \tau_s,$$

for $s = 1, \ldots, s'$. Since $\{y^{*(i)}\}$ is sorted in increasing order, we have $\tau_{s-1} < \tau_s$ for $s = 2, \ldots, s'$. Note that since $u^*(z)$ is strictly increasing when $z \ge 0$ and that $u^*(0) = 0$, Δ_s is a singleton set whenever $\tau_s > 0$. Furthermore, let's denote s^* as the largest index among $1, \ldots, s'$ such that $\tau_{s^*} \le 0$.

Suppose first that $\tau_{s^*} < 0$, let's define a function $\hat{f} : \mathbb{R} \to \mathbb{R}$ in the following way:

$$\hat{f}(\alpha) = \begin{cases} (u^*)^{-1}(\alpha), & \alpha > 0 \\ \mathbf{w}^* \cdot \mathbf{x}^{(k_{s^*})} + \frac{\alpha - \tau_{s^*}}{\tau_{s^*}}(\mathbf{w}^* \cdot \mathbf{x}^{(k_{s^*})}), & \alpha \in [\tau_{s^*}, 0] \\ \mathbf{w}^* \cdot \mathbf{x}^{(k_s)}, & \alpha = \tau_s, s = 1, \ldots, s^* - 1 \\ \mathbf{w}^* \cdot \mathbf{x}^{(k_{s-1})} + \frac{\alpha - \tau_{s-1}}{\tau_s - \tau_{s-1}}(\mathbf{w}^* \cdot \mathbf{x}^{(k_{s-1}+1)} - \mathbf{w}^* \cdot \mathbf{x}^{(k_{s-1})}), & \alpha \in (\tau_{s-1}, \tau_s), s = 2, \ldots, s^* \\ \mathbf{w}^* \cdot \mathbf{x}^{(1)} + \frac{1}{b}(\alpha - \tau_1), & \alpha \in (-\infty, \tau_1) \end{cases} \tag{26}$$

When $\tau_{s^*} = 0$, we define $(0 - \tau_{s^*})/\tau_{s^*} = -1$, and hence $\hat{f}(0) = 0$. The rest remains unchanged. A visualization of $\hat{f}$ with respect to ReLU activation is presented in Figure 3.

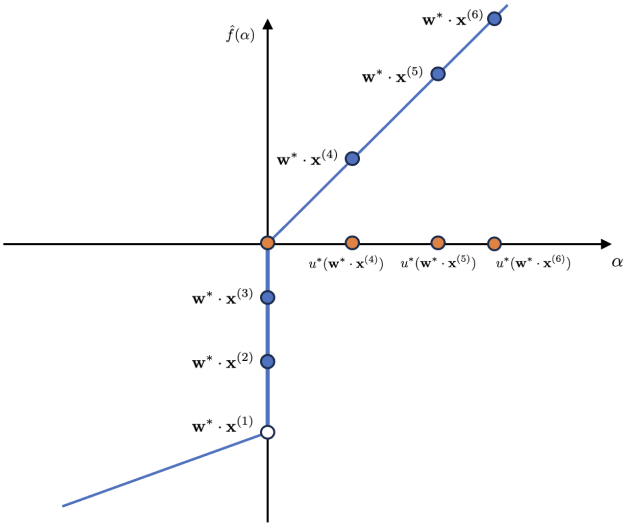

Figure 3: *Given* $u^*(z) = \max\{0, z\}$ *and a dataset* $S^* = \{(\mathbf{x}^{(1)}, u^*(\mathbf{w}^* \cdot \mathbf{x}^{(1)})), \ldots, (\mathbf{x}^{(6)}, u^*(\mathbf{w}^* \cdot \mathbf{x}^{(6)}))\}$ *where* $\mathbf{w}^* \cdot \mathbf{x}^{(1)} < \mathbf{w}^* \cdot \mathbf{x}^{(2)} < \mathbf{w}^* \cdot \mathbf{x}^{(3)} < 0$*,* $\hat{f}$*, the empirical inverse of* u^**, has image above.*

The function $\hat{f}$ has the following properties. First of all, $\hat{f}(\alpha)$ satisfies $\hat{f}(0) = 0$, $(\alpha_1 - \alpha_2)/a \geq \hat{f}(\alpha_1) - \hat{f}(\alpha_2)$, for all $\alpha_1 \geq \alpha_2 \geq 0$, since $\hat{f}(\alpha) = (u^*)^{-1}(\alpha)$ when $\alpha > 0$ and $u^* \in \mathcal{U}_{(a,b)}$. Secondly, $\hat{f}(\alpha_1) - \hat{f}(\alpha_2) \geq (\alpha_1 - \alpha_2)/b$ for all $\alpha_1, \alpha_2 \in \mathbb{R}$, $\alpha_1 \geq \alpha_2$. This is because each segment of $\hat{f}$ has tangent at least $1/b$. Thirdly, for any $\alpha \geq u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$, it holds that $\hat{f}(\alpha) - \mathbf{w}^* \cdot \mathbf{x}^{(i)} \geq (\alpha - u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}))/b$. This is because that suppose $u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}) \in \Delta_s$, for any $\alpha \geq \tau_s$ it holds

$$\hat{f}(\alpha) - \mathbf{w}^* \cdot \mathbf{x}^{(i)} \geq \hat{f}(\alpha) - \mathbf{w}^* \cdot \mathbf{x}^{(k_s)} = \hat{f}(\alpha) - \hat{f}(\tau_s) = \hat{f}(\alpha) - \hat{f}(u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})) \geq (\alpha - u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}))/b,$$

by the fact that tangent of $\hat{f}(\alpha)$ is lower bounded by $1/b$. In addition, for any $\alpha < u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$, it holds $\mathbf{w}^* \cdot \mathbf{x}^{(i)} - \hat{f}(\alpha) \geq (u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}) - \alpha)/b$. This can be seen similarly from the construction of $\hat{f}$. Suppose $u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}) \in \Delta_s$, then for any $\alpha < \tau_s$ it holds

$$\mathbf{w}^* \cdot \mathbf{x}^{(i)} - \hat{f}(\alpha) \geq \mathbf{w}^* \cdot \mathbf{x}^{(k_{s-1}+1)} - \hat{f}(\alpha) = \hat{f}(\tau_s) - \hat{f}(\alpha) = \hat{f}(u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})) - \hat{f}(\alpha) \geq (u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}) - \alpha)/b.$$

Again, we used the fact that $\hat{f}(\alpha_1) - \hat{f}(\alpha_2) \geq (\alpha_1 - \alpha_2)/b$ for all $\alpha_1, \alpha_2 \in \mathbb{R}$, $\alpha_1 \geq \alpha_2$ in the last inequality.

Now we turn to the summation displayed in the statement of the claim. To proceed, we add and subtract $\hat{f}(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}))$ in

the second component in the inner product, which yields:

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\\
&=\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})))\\
&\quad+\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)}). \qquad (27)
\end{aligned}$$

To bound below the first term in (27), we make use of the following Fact C.11. The proof of Fact C.11 can be found at Appendix C.7.

Fact C.11. Let $\mathbf{w}^t\in\mathbb{B}(W)$. Given m samples $S=\{(\mathbf{x}^{(1)},y^{*(1)}),\cdots,(\mathbf{x}^{(m)},y^{*(m)})\}$, let $\hat{u}^{*t}$ be one of the solutions to the optimization problem (P*) , i.e., $\hat{u}^{*t}\in\operatorname{argmin}_{u\in\mathcal{U}_{(a,b)}}(1/m)\sum_{i=1}^{m}(u(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2$. Then

$$\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-f(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})))\geq 0,$$

for any function $f:\mathbb{R}\to\mathbb{R}$ such that $f(0)=0$, $(\alpha_1-\alpha_2)/a\geq f(\alpha_1)-f(\alpha_2)$ for all $\alpha_1\geq\alpha_2\geq 0$, and $f(\alpha_1)-f(\alpha_2)\geq(\alpha_1-\alpha_2)/b$, $\forall\alpha_1,\alpha_2\in\mathbb{R}$, $\alpha_1\geq\alpha_2$.

As we showed previously, $\hat{f}$ satisfies the prerequisite of Fact C.11, hence applying Fact C.11 we observe that the first term in (27) is non-negative:

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})))\geq 0. \qquad (28)$$

Therefore, plugging (28) back into (27), we get:

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\geq\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)}). \qquad (29)$$

Recall that we have showed the function $\hat{f}$ satisfies $\hat{f}(\alpha)-\mathbf{w}^*\cdot\mathbf{x}^{(i)}\geq(\alpha-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))/b\geq 0$ whenever $\alpha\geq u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$, and moreover, $\mathbf{w}^*\cdot\mathbf{x}^{(i)}-\hat{f}(\alpha)\geq(u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})-\alpha)/b\geq 0$ when $\alpha<u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$. Therefore, let $\alpha=\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})$, combining these results we get

$$(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(f(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)})\geq\frac{1}{b}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))^2.$$

Bringing the inequality above back to (29) we then get

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\geq\frac{1}{mb}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2. \qquad (30)$$

The goal now is to bound below the right-hand side of (30) by $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]$ and some small error terms using Chebyshev inequality as we did in Claim C.2. Plugging in Lemma 3.2 we can further lower bound $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]$ by $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$ and then we are done with the proof of this claim. Note that Chebyshev's inequality yields

$$\mathbf{Pr}\left[\left|\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]\right|\geq s\right]\leq\frac{1}{ms^2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^4]. \qquad (31)$$

We now bound $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^4]$. Observe that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^4] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x})+u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2] \\ &\le 4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2((\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2+(y^*)^2)] \\ &\quad + 4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2((\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2+(y^*)^2)]. \end{aligned} \tag{32}$$

We focus on the two terms in (32) separately. Again, choosing $r=\frac{2W}{L}\log(2/(L^2\epsilon'))$, then by the L-sub-exponential tail bound of $\mathcal{D}_\mathbf{x}$, it holds $\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|\ge r]\le\epsilon', \mathbf{Pr}[|\mathbf{w}^*\cdot\mathbf{x}|\ge r]\le\epsilon'$. Since $y^*=u^*(\mathbf{w}^*\cdot\mathbf{x})$ and both u^* and $\hat{u}^{*t}$ are non-decreasing b-Lipschitz, it holds:

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2((\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2+(y^*)^2)] \\ &\le b^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2((\mathbf{w}^t\cdot\mathbf{x})^2+(\mathbf{w}^*\cdot\mathbf{x})^2)] \\ &= b^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2((\mathbf{w}^t\cdot\mathbf{x})^2+(\mathbf{w}^*\cdot\mathbf{x})^2)\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\le r,|\mathbf{w}^*\cdot\mathbf{x}|\le r\}] \\ &\quad + b^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2((\mathbf{w}^t\cdot\mathbf{x})^2+(\mathbf{w}^*\cdot\mathbf{x})^2)\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\ge r \text{ or } |\mathbf{w}^*\cdot\mathbf{x}|\ge r\}] \\ &\le 2b^2r^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] \\ &\quad + 2b^4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[2(\mathbf{w}^t\cdot\mathbf{x})^2((\mathbf{w}^t\cdot\mathbf{x})^2+(\mathbf{w}^*\cdot\mathbf{x})^2)\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\ge r \text{ or } |\mathbf{w}^*\cdot\mathbf{x}|\ge r\}]. \end{aligned} \tag{33}$$

The first term in (33) can be upper bounded using Lemma F.2, which states that when $m \gtrsim d\log(1/\delta)(b^2W^3\log^2(d/\epsilon)/(L^2\epsilon))^{3/2})$, with probability at least $1-\delta$ it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]\le\epsilon$ for all $\mathbf{w}^t\in\mathbb{B}(W)$. Now suppose this inequality is valid the given $\mathbf{w}^t\in\mathbb{B}(W)$ (which happens with probability at least $1-\delta$). For the second term in (33), note that for any unit vector $\mathbf{a}$ it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{a}\cdot\mathbf{x})^8]\le c^2/L^8$ for some absolute constant $c>0$, and furthermore, the magnitude of r ensures that $\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|\ge r \text{ or } |\mathbf{w}^*\cdot\mathbf{x}|\ge r]\le 2\epsilon'$, therefore, combining these bounds, we get:

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[2(\mathbf{w}^t\cdot\mathbf{x})^2((\mathbf{w}^t\cdot\mathbf{x})^2+(\mathbf{w}^*\cdot\mathbf{x})^2)\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\ge r \text{ or } |\mathbf{w}^*\cdot\mathbf{x}|\ge r\}] \\ &\le 2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}^t\cdot\mathbf{x})^8]\,\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|\ge r \text{ or } |\mathbf{w}^*\cdot\mathbf{x}|\ge r]} \\ &\quad + 2\sqrt{2(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}^t\cdot\mathbf{x})^8]+\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}^*\cdot\mathbf{x})^8])\,\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|\ge r \text{ or } |\mathbf{w}^*\cdot\mathbf{x}|\ge r]} \\ &\le 24c(W/L)^4\sqrt{\epsilon'}. \end{aligned}$$

Plugging back into (33), we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2((\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2+(y^*)^2)]\le 2b^2r^2\epsilon+48c(bW/L)^4\sqrt{\epsilon'},$$

which is the upper bound on the first term of (32).

For the second term in (32), since by definition we have $u^{*t}\in\operatorname{argmin}_{u\in\mathcal{U}_{(a,b)}}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]$, it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]\le\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^*(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\le b^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x})^2]\le b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2,$$

since $\mathbf{x}$ is isotropic. Thus, using similar steps as in (33), we have

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2((\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2+(y^*)^2)] \\ &\le 2b^2r^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2] \\ &\quad + 2b^4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[2((\mathbf{w}^t\cdot\mathbf{x})^2+(\mathbf{w}^*\cdot\mathbf{x})^2)^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\ge r \text{ or } |\mathbf{w}^*\cdot\mathbf{x}|\ge r\}] \\ &\le 2b^4r^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+48c(bW/L)^4\sqrt{\epsilon}. \end{aligned}$$

In summary, combining all the results and plugging them back into (32), we finally get the upper bound for the variance:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^4]\le\frac{32b^2W^2}{L^2}\log^2(2/(L^2\epsilon'))(b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+\epsilon)+384c(bW/L)^4\sqrt{\epsilon'}.$$

Let $s=b\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2+\epsilon/b$ and plug the last inequality back into (31) to get:

$$\begin{aligned}\mathbf{Pr}\left[\left|\frac{1}{m}\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]\right|\ge b\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2+\epsilon/b\right]\\ \le\frac{1}{m(\epsilon b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+\epsilon^2/b^2)}\left(\frac{32b^2W^2}{L^2}\log^2\left(\frac{2}{L^2\epsilon'}\right)(b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+\epsilon)+384c(bW/L)^4\sqrt{\epsilon'}\right).\end{aligned}$$

Choosing $\epsilon'=\epsilon^2/b^4$ and using similar arguments as in Claim C.2, we get that the right-hand side of the inequality above is bounded by δ, given our choice of $m\gtrsim db^4W^{9/2}\log^4(d/(\epsilon\delta))(1/\epsilon^{3/2}+1/(\epsilon\delta))$ as specified in the statement of Proposition 3.1. In summary, after a union bound on the probability above and the event that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]\le\epsilon$, we have with probability at least $1-2\delta$,

$$\frac{1}{m}\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2\ge\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]-\sqrt{\epsilon}b\|\mathbf{w}^t-\mathbf{w}^*\|_2-\epsilon/b.$$

Recall that in Lemma 3.2 we showed that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\ge Ca^2LR^4\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$ for an absolute constant C; thus, our final result is that with probability at least $1-\delta$,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}&\ge\frac{1}{mb}\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2\\ &\ge\frac{Ca^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2-\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2-\epsilon/b.\end{aligned}$$

This completes the proof of Claim C.3. □

C.7. Proof of Fact C.11

We prove a modified version of Lemma 1 (Kakade et al., 2011), presented as the statement below. The statement considers a smaller activation class and a function f with different properties compared to (Kakade et al., 2011), and the proof is based on a rigorous KKT argument.

Fact C.12. Let $\mathbf{w}^t\in\mathbb{B}(W)$. Given m samples $S=\{(\mathbf{x}^{(1)},y^{*(1)}),\cdots,(\mathbf{x}^{(m)},y^{*(m)})\}$, let $\hat{u}^{*t}$ be one of the solutions to the optimization problem (P*) , i.e., $\hat{u}^{*t}\in\operatorname{argmin}_{u\in\mathcal{U}_{(a,b)}}(1/m)\sum_{i=1}^m(u(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2$. Then

$$\sum_{i=1}^m(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-f(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})))\ge0,$$

for any function $f:\mathbb{R}\to\mathbb{R}$ such that $f(0)=0$, $(\alpha_1-\alpha_2)/a\ge f(\alpha_1)-f(\alpha_2)$ for all $\alpha_1\ge\alpha_2\ge0$, and $f(\alpha_1)-f(\alpha_2)\ge(\alpha_1-\alpha_2)/b$, $\forall\alpha_1,\alpha_2\in\mathbb{R}$, $\alpha_1\ge\alpha_2$.

Proof. We transform the optimization problem (P*) to a quadratic optimization problem with linear constraints. To guarantee that the solution of this quadratic problem corresponds to a function that is (a,b)-unbounded, we add a sample $(\mathbf{x}^{(k)},y^{*(k)})=(\mathbf{0},0)$ to the sample set. Let $z_i=\mathbf{w}^t\cdot\mathbf{x}^{(i)}$ such that (perhaps after some permutation) $z_1\le z_2\le\cdots\le z_m$ and $z_k=0$, we solve the following optimization problem:

$$\begin{aligned}\min_{\tilde{y}^{(i)},i\in[m]}\ &\sum_{i=1}^m(\tilde{y}^{(i)}-y^{*(i)})^2\\ \text{s.t. }&0\le\tilde{y}^{(i+1)}-\tilde{y}^{(i)},\ 1\le i\le k-1\ ,\\ &a(z_{i+1}-z_i)\le\tilde{y}^{(i+1)}-\tilde{y}^{(i)},\ k\le i\le m-1\ ,\\ &\tilde{y}^{(i+1)}-\tilde{y}^{(i)}\le b(z_{i+1}-z_i),\ 1\le i\le m-1\ ,\\ &\tilde{y}^{(k)}=0\ .\end{aligned}\tag{34}$$

Denote the solution of (34) as $\hat{y}^{*(i)}$, $i = 1, \cdots, m$. Let $\hat{u}^{*t}(z)$ be the linear interpolation function of $(z_i, \hat{y}^{*(i)})$, then $\hat{u}^{*t} \in \mathcal{U}_{(a,b)}$ since $\hat{u}^{*t}(0) = \hat{u}^{*t}(z_k) = \hat{y}^{*(k)} = 0$, $\hat{u}^{*t}$ is b-Lipschitz and $\hat{u}^{*t}(z) - \hat{u}^{*t}(z') \geq a(z - z')$ for all $z \geq z' \geq 0$. In other words, finding a solution of (P*) is equivalent to solving (34).

Now observe that the summation $\sum_{i=1}^m (\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)}))$ can be transformed into the following:

$$\sum_{i=1}^m (\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)})) = \sum_{i=1}^m \left(\sum_{j=1}^i (\hat{y}^{*(j)} - y^{*(j)}) \right) (z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))), \tag{35}$$

where we let $z_{m+1} = 0$, $\hat{y}^*_{m+1} = 0$ (and hence $f(\hat{y}^*_{m+1}) = 0$ as $f(0) = 0$).

To fully utilize the information that $\hat{y}^{*(i)}$ is the minimizer of the optimization problem (34), we write down the KKT condition for the optimization problem (34) described above:

$$\hat{y}^{*(i)} = y^{*(i)} + (\lambda'_i - \lambda'_{i-1})/2 - (\lambda_i - \lambda_{i-1})/2 - (\nu_k/2)\mathbb{1}\{i = k\},\ i = 1, \cdots, m; \tag{36}$$

$$-\lambda_i(\hat{y}^{*(i+1)} - \hat{y}^{*(i)}) = 0,\ i = 1, \cdots, k-1; \tag{37}$$

$$\lambda_i(a(z_{i+1} - z_i) - (\hat{y}^{*(i+1)} - \hat{y}^{*(i)})) = 0,\ i = k, \cdots, m-1; \tag{38}$$

$$\lambda'_i((\hat{y}^{*(i+1)} - \hat{y}^{*(i)}) - b(z_{i+1} - z_i)) = 0,\ i = 1, \cdots, m-1; \tag{39}$$

$$\nu_k \hat{y}^{*(k)} = 0\ , \tag{40}$$

where $\lambda_i, \lambda'_i \geq 0$, for $i = 1, \ldots, m-1$, and $\nu_k \in \mathbb{R}$ are dual variables, and we let $\lambda_0 = \lambda'_0 = 0$ for the convenience of presenting (36).

Summing up (36) recursively, we immediately get that

$$\sum_{j=1}^i (\hat{y}^{*(i)} - y^{*(i)}) = \frac{1}{2}((\lambda'_i - \lambda_i) - \nu_k \mathbb{1}\{i \geq k\}).$$

Bringing this equality above back to (35), we have

$$\begin{aligned}
&\sum_{i=1}^m (\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)})) \\
&= \frac{1}{2}\sum_{i=1}^m (\lambda'_i - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) + \frac{1}{2}\sum_{i=k}^m \nu_k (z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \\
&= \frac{1}{2}\sum_{i=1}^m (\lambda'_i - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) + \nu_k (z_k - f(\hat{y}^{*(k)}) - (z_{m+1} - f(\hat{y}^{*(m+1)}))).
\end{aligned} \tag{41}$$

Since by definition, $z_{m+1} = f(\hat{y}^{*(m+1)}) = 0$, $z_k = 0$, and as $\hat{y}^{*(i)}$, $i \in [m]$, is a feasible solution of (34), it holds $\hat{y}^{*(k)} = 0$, we thus have

$$\nu_k (z_k - f(\hat{y}^{*(k)}) - (z_{m+1} - f(\hat{y}^{*(m+1)}))) = 0.$$

Bringing this back to (41), we get

$$\sum_{i=1}^m (\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)})) = \frac{1}{2}\sum_{i=1}^m (\lambda'_i - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))). \tag{42}$$

Consider first when $i = 1, \ldots, k-1$. Suppose for some $i \in \{1, \ldots, k-1\}$ we have $\lambda'_i, \lambda_i > 0$, then according to the complementary slackness condition (37) and (38) it holds that $0 = \hat{y}^{*(i+1)} - \hat{y}^{*(i)} = b(z_{i+1} - z_i)$. Therefore, the we have $(\lambda'_i - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)})))$. Suppose for some $i \in \{1, \ldots, k-1\}$, it holds $\lambda'_i > 0$. Then, it must be the case that $\hat{y}^{*(i+1)} - \hat{y}^{*(i)} = b(z_{i+1} - z_i) \geq 0$, according to the KKT condition (39). Since $f(\hat{y}^{*(i+1)}) - f(\hat{y}^{*(i)}) \geq (\hat{y}^{*(i+1)} - \hat{y}^{*(i)})/b$ by assumption on f, we thus have $(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \geq 0$. Finally, if $\lambda_i > 0$, then (37) indicates that $0 = \hat{y}^{*(i+1)} - \hat{y}^{*(i)}$. Therefore, as $z_{i+1} \geq z_i$, the i^{th} summand is also positive. In summary, the first summation in (42) is positive.

Now consider $i \in \{k, \dots, m\}$. Observe that if for some $i \in \{k, \dots, m\}$ it holds $\lambda_i > 0$ and $\lambda_i' > 0$ at the same time, then KKT conditions (38) and (39) imply that $a(z_{i+1} - z_i) = \hat{y}^{*(i+1)} - \hat{y}^{*(i)} = b(z_{i+1} - z_i)$, as $a < b$ it has to be $z_{i+1} - z_i = \hat{y}^{*(i+1)} - \hat{y}^{*(i)} = 0$, which indicates that the i^{th} summand in the second term must be 0, i.e., $(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) = 0$.

Now suppose for some $i \in \{1, \dots, m\}$, $\lambda_i' > 0$ and $\lambda_i = 0$, then by the complementary slackness conditions (38) and (39), it must be that $\hat{y}^{*(i+1)} - \hat{y}^{*(i)} = b(z_{i+1} - z_i) \geq 0$. Again, since f satisfies $f(\hat{y}^{*(i+1)}) - f(\hat{y}^{*(i)}) \geq (\hat{y}^{*(i+1)} - \hat{y}^{*(i)})/b$ for any $\hat{y}^{*(i+1)} \geq \hat{y}^{*(i)}$, we thus have $z_i - z_{i+1} + (f(\hat{y}^{*(i+1)}) - f(\hat{y}^{*(i)})) \geq 0$. Thus it holds that $(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \geq 0$.

On the other hand, if $\lambda_i' = 0$ and $\lambda_i > 0$, then complementary slackness implies that $\hat{y}^{*(i+1)} - \hat{y}^{*(i)} = a(z_{i+1} - z_i) \geq 0$. Furthermore, since $\hat{y}^{*(i)} \geq \hat{y}^{*(k)} \geq 0$ when $i \geq k$, using the assumption that $(\alpha_1 - \alpha_2)/a \geq f(\alpha_1) - f(\alpha_2)$ when $\alpha_1 \geq \alpha_2 \geq 0$, we get $z_i - z_{i+1} + (f(\hat{y}^{*(i+1)}) - f(\hat{y}^{*(i)})) \leq 0$, and hence $(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \geq 0$ holds as well.

In summary, since each summand in (42) is positive, we finally get that

$$\sum_{i=1}^{m} (\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)})) = \sum_{i=1}^{m} \left(\sum_{j=1}^{i} (\hat{y}^{*(j)} - y^{*(j)}) \right) (z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \geq 0.$$

This completes the proof of Fact C.11. □

C.8. Proof of Claim C.4

We restate and prove Claim C.4 that appeared in the proof of Proposition 3.1 in Appendix C.1.

Claim C.13. Let $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ be m i.i.d. samples from $\mathcal{D}$, and denote $S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}_{i=1}^m$ the uncorrupted version of S where $y^{*(i)} = u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$. Under the condition of Proposition 3.1, given a parameter $\mathbf{w}^t \in \mathbb{B}(W)$ with probability at least $1 - \delta$ it holds:

$$\frac{1}{m} \sum_{i=1}^{m} (y^{*(i)} - y^{(i)})(\mathbf{w}^t \cdot \mathbf{x}^{(i)} - \mathbf{w}^* \cdot \mathbf{x}^{(i)}) \geq -\sqrt{\mathrm{OPT}} \|\mathbf{w}^* - \mathbf{w}^t\|_2 - (\mathrm{OPT} + \epsilon)/b \,.$$

Proof. By Chebyshev's inequality, we can write

$$\mathbf{Pr}\left[\left| \frac{1}{m} \sum_{i=1}^{m} (y^{*(i)} - y^{(i)})(\mathbf{w}^t \cdot \mathbf{x}^{(i)} - \mathbf{w}^* \cdot \mathbf{x}^{(i)}) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [(y^* - y)(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}] \right| \geq s \right] \leq \frac{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}} [(y^* - y)^2 (\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2]}{ms^2}.$$

Let $r = \frac{2W}{L} \log(2/(L^2\epsilon'))$, then by the fact that $\mathcal{D}_{\mathbf{x}}$ is sub-exponential, we have $\mathbf{Pr}[|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \geq r] \leq \epsilon'$. Furthermore, since $|y| \leq M$ where $M = \frac{bW}{L} \log(16b^4W^4/\epsilon^2)$, as stated in Fact F.3, the variance can be bounded as follows:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [(y^* - y)^2 (\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2] \\
&\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [(y^* - y)^2 (\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \leq r\}] \\
&\quad + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [(y^* - y)^2 (\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \geq r\}] \\
&\leq r^2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [(u^*(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] \\
&\quad + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [(2(u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 + y^2)(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \geq r\}] \\
&\leq r^2 \mathrm{OPT} + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}} [2(b^2(\mathbf{w}^t \cdot \mathbf{x})^2 + M^2)(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \geq r\}].
\end{aligned}$$

Since for any unit vectors $\mathbf{a}, \mathbf{b}$ we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{a}\cdot\mathbf{x})^4] \le c^2/L^4$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{a}\cdot\mathbf{x})^4(\mathbf{b}\cdot\mathbf{x})^4] \le c^2/L^8$, we have:

$$\begin{aligned}
&2b^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}^t\cdot\mathbf{x})^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x}^2)^2\mathbb{1}\{|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\ge r\}]\\
&\le 4b^2(W/L)^4\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t/\|\mathbf{w}^t\|_2)\cdot\mathbf{x})^4(((\mathbf{w}^t-\mathbf{w}^*)/\|\mathbf{w}^t-\mathbf{w}^*\|_2)\cdot\mathbf{x})^4]\,\mathbf{Pr}[|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\ge r]}\\
&\le 4cb^2(W/L)^4\sqrt{\epsilon'},
\end{aligned}$$

and in addition,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[M^2((\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x})^2\mathbb{1}\{|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\ge r\}]\\
&\le 2M^2W^2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x})^4]\,\mathbf{Pr}[|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\ge r]} \le cM^2(W/L)^2\sqrt{\epsilon'}.
\end{aligned}$$

Let $s = (\mathrm{OPT}+\epsilon)/b$, $\epsilon' = \epsilon^2$, under our choice of $m \gtrsim db^4W^{9/2}\log^4(d/(\epsilon\delta))(1/\epsilon^{3/2}+1/(\epsilon\delta))$, it holds that

$$\frac{1}{ms^2}\left(\frac{4W^2\log^2(1/(L^2\epsilon'))\mathrm{OPT}}{L^2} + (4cb^2(W/L)^4 + cM^2(W/L)^2)\sqrt{\epsilon'}\right) \le \delta.$$

Thus, with probability at least $1-\delta$ it holds that

$$\frac{1}{m}\sum_{i=1}^m (y^{*(i)}-y^{(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\mathbf{w}^*\cdot\mathbf{x}^{(i)}) \ge \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-y^*)(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}] - (\mathrm{OPT}+\epsilon)/b.$$

Since

$$\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-y^*)(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}]\right| \le \sqrt{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-y^*)^2]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x})^2]} \le \sqrt{\mathrm{OPT}}\|\mathbf{w}^*-\mathbf{w}^t\|_2,$$

we finally have

$$\frac{1}{m}\sum_{i=1}^m (y^{*(i)}-y^{(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\mathbf{w}^*\cdot\mathbf{x}^{(i)}) \ge -\sqrt{\mathrm{OPT}}\|\mathbf{w}^*-\mathbf{w}^t\|_2 - (\mathrm{OPT}+\epsilon)/b,$$

completing the proof of Claim C.4. □

D. Omitted Proofs from Section 4

D.1. Proof of Lemma 4.1

Our algorithm for initialization is the following Algorithm 2:

Algorithm 2 Initialization

1: **Input:** $\mathbf{w}^0 = 0$; $\epsilon, \delta > 0$; positive parameters a, b, L, R, W; $\mu \lesssim a^2LR^4/b$, step size $\eta = \mu^3/(2^7b^4)$, number of iterations $t_0 \lesssim (b/\mu)^6\log(b/\mu)$;
2: **for** $t = 0$ **to** t_0 **do**
3: Draw $m_0 \gtrsim W^{9/2}b^{10}d\log^4(d/(\epsilon\delta))/(L^4\mu^6\delta\epsilon^{3/2})$ i.i.d. samples from $\mathcal{D}$
4: $\hat{u}^t = \mathop{\mathrm{argmin}}_{u\in\mathcal{U}_{(a,b)}} \frac{1}{m_0}\sum_{i=1}^{m_0}(u(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{(i)})^2$.
5: $\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t) = \frac{1}{m_0}\sum_{i=1}^{m_0}(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{(i)})\mathbf{x}^{(i)}$.
6: $\mathbf{w}^{t+1} = \mathbf{w}^t - \eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)$.
7: **end for**
8: **Return:** $\{\mathbf{w}^0,\dots,\mathbf{w}^{t_0}\}$

We restate and prove Lemma 4.1 in below.

Lemma D.1 (Initialization). *Let $\mu = Ca^2LR^4/b$ for an absolute constant $C > 0$ and let $\epsilon, \delta > 0$. Choose the step size $\eta = \mu^3/(2^7 b^4)$ in Algorithm 2. Then, drawing m_0 i.i.d. samples from $\mathcal{D}$ at each iteration such that*

$$m_0 \gtrsim \frac{W^{9/2} b^{10} d \log^4(d/(\epsilon\delta))}{L^4 \mu^6 \delta \epsilon^{3/2}},$$

ensures that within $t_0 \lesssim b^6 \log(b/\mu)/\mu^6$ iterations, the initialization subroutine Algorithm 2 generates a list of size t_0 that contains a point $\bar{\mathbf{w}}^0$ such that $\|(\mathbf{w}^*)^{\perp_{\bar{\mathbf{w}}^0}}\|_2 \le \max\{\mu\|\mathbf{w}^*\|_2/(4b), 64b^2/\mu^3(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\}$, with probability at least $1-\delta$. The total number of samples required for Algorithm 2 is $N_0 = t_0 m_0$.

Proof. Assume first that $\|\mathbf{w}^*\|_2 \le 64b^2/\mu^3(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$. Then, for the parameter $\mathbf{w}^0 = 0$, it holds that $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^0}}\|_2 = \|\mathbf{w}^*\|_2 \le 64b^2/\mu^3(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$. Hence, $\mathbf{w}^0 = 0$ satisfies the condition $\|(\mathbf{w}^*)^{\perp_{\bar{\mathbf{w}}^0}}\|_2 \le \max\{\mu\|\mathbf{w}^*\|_2/(4b), 64b^2/\mu^3(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\}$.

Now suppose $\|\mathbf{w}^*\|_2 \ge 64b^2/\mu^3(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$. Let $\mathbf{v}^t$ denote the component of $\mathbf{w}^*$ that is orthogonal to $\mathbf{w}^t$; i.e., $\mathbf{v}^t = \mathbf{w}^* - (\mathbf{w}^* \cdot \mathbf{w}^t)\mathbf{w}^t/\|\mathbf{w}^t\|_2^2 = (\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}$, where $\mathbf{w}^t$ is defined in Algorithm 2. Our goal is to show that when $\|\mathbf{v}^t\|_2 \ge \mu\|\mathbf{w}^*\|_2/(4b)$ at iteration t, the distance between $\mathbf{w}^{t+1}$ and $\mathbf{w}^*$ contracts by a constant factor $1-c$ for some $c < 1$, i.e., $\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \le (1-c)\|\mathbf{w}^t - \mathbf{w}^*\|_2$. This implies that when $\|\mathbf{v}^t\|_2$ is greater than $\mu\|\mathbf{w}^*\|_2/(4b)$, $\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_2$ contracts until $\|\mathbf{v}^t\|_2 \ge \mu\|\mathbf{w}^*\|_2/(4b)$ is violated at step t_0; this $\mathbf{w}^{t_0}$ is exactly the initial point we are seeking to initialize the optimization subroutine.

Applying Proposition 3.1 we get that under our choice of batch size m, with probability at least $1-\delta$, at each iteration it holds

$$\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) \ge \frac{Ca^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2 - 2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2 - 2(\mathrm{OPT} + \epsilon)/b.$$

We now study the distance between $\mathbf{w}^{t+1}$ and $\mathbf{w}^*$, where $\mathbf{w}^{t+1}$ is updated from $\mathbf{w}^t$ according to Algorithm 2.

$$\begin{aligned}\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 &= \|\mathbf{w}^t - \eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) - \mathbf{w}^*\|_2^2 \\ &= \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \eta^2\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2^2 - 2\eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*). \end{aligned} \tag{43}$$

Applying Lemma 4.3 to (43), and plugging in Proposition 3.1, we get that under our choice of batch size m it holds that with probability at least $1-\delta$,

$$\begin{aligned}\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 &\le \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \eta^2(10(\mathrm{OPT} + \epsilon) + 4b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2) \\ &\quad + 2\eta(2(\mathrm{OPT} + \epsilon)/b + 2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2 - \mu\|\mathbf{v}^t\|_2^2) \\ &\le (1 + 4b^2\eta^2)\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 2\eta(2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2 - \mu\|\mathbf{v}^t\|_2^2) \\ &\quad + 5\eta(\mathrm{OPT} + \epsilon), \end{aligned} \tag{44}$$

where $\mu = Ca^2LR^4/b$ and C is an absolute constant. Note that in the last inequality we used that $\eta \le 1/10$, hence $10\eta^2 \le \eta$, and that $b \ge 1$.

Note that we have assumed $\sqrt{\mathrm{OPT}} + \sqrt{\epsilon} \le \mu^3/(64b^2)\|\mathbf{w}^*\|_2$. Furthermore, when $t = 0$, $\mathbf{v}^0 = \mathbf{w}^*$ hence we would have $\|\mathbf{v}^0\|_2 \ge \mu\|\mathbf{w}^*\|_2/(4b)$. Suppose that at iteration t, $\|\mathbf{v}^t\|_2 \ge \mu\|\mathbf{w}^*\|_2/(4b)$ is still valid. Then, (44) is transformed to:

$$\begin{aligned}\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 &\le (1 + 4b^2\eta^2)\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 5\eta(\mathrm{OPT} + \epsilon) \\ &\quad + 2\eta((\mu^3/(32b^2))\|\mathbf{w}^t - \mathbf{w}^*\|_2\|\mathbf{w}^*\|_2 - (\mu^3/(16b^2))\|\mathbf{w}^*\|_2^2) \end{aligned} \tag{45}$$

We will use an induction argument to show that at iteration t, $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \le \|\mathbf{w}^*\|_2$, which will eventually yield a contraction $\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 \le (1-c)\|\mathbf{w}^t - \mathbf{w}^*\|_2^2$ for some constant $c < 1$. This condition $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \le \|\mathbf{w}^*\|_2$ certainly holds for the base case $t = 0$ as $\mathbf{w}^0 = 0$ hence $\|\mathbf{w}^0 - \mathbf{w}^*\|_2 = \|\mathbf{w}^*\|_2$. Now, suppose $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \le \|\mathbf{w}^*\|_2$ holds for all the iterations from 0 to t. Then, bringing in $\eta = \mu^3/(2^7 b^4)$ to (45), we get:

$$\begin{aligned}\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 &\le (1 + 4b^2\eta^2)\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 2\eta((\mu^3/(32b^2)) - (\mu^3/(16b^2)))\|\mathbf{w}^t - \mathbf{w}^*\|_2\|\mathbf{w}^*\|_2 + 5\eta(\mathrm{OPT} + \epsilon) \\ &\le (1 + 4\eta^2b^2 - 2\eta\mu^3/(32b^2))\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 5\mu^3/(2^7b^4)(\mathrm{OPT} + \epsilon) \\ &\le (1 - \mu^6/(2^{11}b^6))\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 5\mu^3/(2^7b^4)(\mathrm{OPT} + \epsilon). \end{aligned}$$

Since we have assumed $\sqrt{\mathrm{OPT}}+\sqrt{\epsilon} \leq \mu^3/(64b^2)\|\mathbf{w}^*\|_2$, it holds $\|\mathbf{w}^t-\mathbf{w}^*\|_2 \geq \|\mathbf{v}^t\|_2 \geq \mu\|\mathbf{w}^*\|_2/(4b) \geq (16b/\mu^2)(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})$, thus, we have (noting that $\mu \leq 1$):

$$5\mu^3/(2^7b^4)(\mathrm{OPT}+\epsilon) \leq 5\mu^3/(2^7b^4)(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})^2 \leq \mu^6/(2^{12}b^6)\|\mathbf{w}^t-\mathbf{w}^*\|_2.$$

Therefore, combining the results above, we get:

$$\|\mathbf{w}^{t+1}-\mathbf{w}^*\|_2^2 \leq (1-\mu^6/(2^{12}b^6))\|\mathbf{w}^t-\mathbf{w}^*\|_2^2,$$

for any iteration t such that $\|\mathbf{v}^t\|_2 \geq \mu\|\mathbf{w}^*\|_2/(4b)$ holds. This validates the induction argument that $\|\mathbf{w}^t-\mathbf{w}^*\|_2 \leq \|\mathbf{w}^*\|_2$ for every $t=0,\ldots,t_0$ and at the same time yields the desired contraction property of the sequence $\|\mathbf{w}^t-\mathbf{w}^*\|_2$, $t=0,\ldots,t_0$. Now, since $\|\mathbf{w}^0-\mathbf{w}^*\|_2=\|\mathbf{w}^*\|_2$ and $\|\mathbf{w}^t-\mathbf{w}^*\|_2 \geq \|\mathbf{v}^t\|_2$, we have

$$\|\mathbf{v}^{t+1}\|_2^2 \leq (1-\mu^6/(2^{12}b^6))^t\|\mathbf{w}^*\|_2^2 \leq \exp(-t\mu^6/(2^{12}b^6))\|\mathbf{w}^*\|_2^2.$$

Thus, after at most $t_0=2^{12}b^6\log(4b/\mu)/\mu^6$ iterations, it must hold that among all those vectors $\mathbf{v}^1,\ldots,\mathbf{v}^{t_0}$, there exists a vector $\mathbf{v}^{t_0^*}$ such that $\|\mathbf{v}^{t_0^*}\|_2 \leq \mu\|\mathbf{w}^*\|_2/(4b)$. Since there are only a constant number of candidates, we can feed each one as the initialized input to the optimization subroutine Algorithm 3. This will only result in a constant boost up in the runtime and sample complexity.

Finally, recall that we need to draw

$$m \gtrsim \frac{W^{9/2}b^4\log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}}+\frac{1}{\epsilon\delta}\right).$$

new samples at each iteration for (44) to hold with probability $1-\delta$, and the total number of iterations is t_0. Thus, applying a union bound we know that the probability that (44) holds for all t_0 is $1-t_0\delta$. Hence, choosing $\delta \leftarrow \delta t_0$, and note that $t_0 \approx b^6/\mu^6\log(b/\mu)$, it yields that setting the batch size to be

$$m_0=\Theta\left(\frac{W^{9/2}b^4\log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}}+\frac{b^6\log(b/\mu)}{\mu^6\epsilon\delta}\right)\right)=\Theta\left(\frac{W^{9/2}b^{10}d\log^4(d/(\epsilon\delta))}{L^4\mu^6\delta\epsilon^{3/2}}\right),$$

suffices and the total number of sample complexity for the initialization process is t_0m_0. □

D.2. Proof of Theorem 4.2

In this subsection, we restate and prove our main theorem Theorem 4.2. The full version of the optimization algorithm as well as the main theorem Theorem 4.2 is displayed below:

Theorem D.2 (Main Result). *Let $\mathcal{D}$ be a distribution in $\mathbb{R}^d\times\mathbb{R}$ and suppose that $\mathcal{D}_{\mathbf{x}}$ is (L,R)-well-behaved. Furthermore, let $\mathcal{U}_{(a,b)}$ be as in Definition 1.3, and $\epsilon>0$. Let $\mu=Ca^2LR^4/b$, where C is an absolute constant. Running Algorithm 1 with the following parameters: step size $\eta=\mu/(4b^2)$, batch size to be $m \gtrsim dW^{5.5}b^{14}\log^5(d/\epsilon)/(L^4\mu^9\epsilon^{3/2})$ and the total number of iterations to be $T'=t_0JT=O(Wb^{10}/(\mu^9\sqrt{\epsilon})\log(1/\epsilon))$, where $T=O((b/\mu)^2\log(1/\epsilon))$, then with probability at least $2/3$, Algorithm 1 returns a hypothesis $(\hat{u},\widehat{\mathbf{w}})$ where $\hat{u}\in\mathcal{U}_{(a,b)}$ and $\widehat{\mathbf{w}}\in\mathbb{B}(W)$ such that*

$$\mathcal{L}_2(\widehat{\mathbf{w}};\hat{u})=O\left(\frac{b^4}{a^4L^2R^8}\right)\mathrm{OPT}+\epsilon\,,$$

using $N=O(T'm)=\tilde{O}(dW^{6.5}b^{24}/(L^4\mu^{18}\epsilon^2))$ samples.

Proof. As proved in Lemma 4.1, the initialization subroutine Algorithm 2 outputs a starting point $\bar{\mathbf{w}}^0$ such that $\|(\mathbf{w}^*)^{\perp_{\bar{\mathbf{w}}^0}}\|_2 \leq \max\{\mu\|\mathbf{w}^*\|_2/(4b), 64b^2/\mu^3(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})\}$. Suppose first that $\mu\|\mathbf{w}^*\|_2/(4b) \leq 64b^2/\mu^3(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})$. Then this implies that $\|\mathbf{w}^*\|_2 \leq 256b^3/\mu^4(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})$. Therefore, applying Claim 4.4 we immediately get that the trivial hypothesis $(\mathbf{w}=0,u(z)=0)$ works as a constant approximate solution, as $\mathcal{L}_2(\mathbf{w};u) \leq 8(\mathrm{OPT}+\epsilon)+4b^2\|\mathbf{w}^*\|_2 = O((b/\mu)^8)\mathrm{OPT}+\epsilon$. This hypothesis $(\mathbf{w}=0,u(z)=0)$ is contained in our solution set $\mathcal{P}$ (see Algorithm 3) and will be tested in Algorithm 4. Thus, we assume this is not the case and the initial point $\bar{\mathbf{w}}^0$ satisfies $\|(\mathbf{w}^*)^{\perp_{\bar{\mathbf{w}}^0}}\|_2 \leq \mu\|\mathbf{w}^*\|_2/(4b)$.

Since there exists a $\mathbf{w}_{k^*}^{\mathrm{ini}} \in \{\mathbf{w}_k^{\mathrm{ini}}\}_{k=1}^{t_0}$ such that $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}_{k^*}^{\mathrm{ini}}}}\|_2 \leq \mu\|\mathbf{w}^*\|_2/(4b)$. Let us consider this initialized parameter at k^* step $\bar{\mathbf{w}}_{j,k^*}^0=\mathbf{w}_{k^*}^{\mathrm{ini}}$ and ignore the subscript k^* for simplicity. Since we constructed a grid with grid width $\eta\sqrt{\epsilon}$ from

Algorithm 3 Optimization

Input: $\mathbf{w}^{\mathrm{ini}} = \mathbf{0}$; $\epsilon > 0$; positive parameters: a, b, L, R, W; let $\mu \lesssim a^2LR^4/b$; step size $\eta = \mu/(4b^2)$, number of iterations $T = O((b/\mu)^2 \log(1/\epsilon))$.
$\{\mathbf{w}_0^{\mathrm{ini}}, \dots, \mathbf{w}_{t_0}^{\mathrm{ini}}\} = \text{Initialization}[\mathbf{w}^{\mathrm{ini}}]$ (Algorithm 2)
for $k = 0$ **to** $t_0 \lesssim (b/\mu)^6 \log(b/\mu)$ **do**
  $\mathcal{P}_k = \{\}$
  for $j = 1$ **to** $J = W/(\eta\sqrt{\epsilon})$ **do**
    $\bar{\mathbf{w}}_{j,k}^0 = \mathbf{w}_k^{\mathrm{ini}}$.
    $\beta_j = j\eta\sqrt{\epsilon}$. ▷ find an $\eta\sqrt{\epsilon}$ approximation of $\|\mathbf{w}^*\|_2$
    for $t = 0$ **to** $T - 1$ **do**
      $\widehat{\mathbf{w}}_{j,k}^t = \beta_j(\bar{\mathbf{w}}_{j,k}^t/\|\bar{\mathbf{w}}_{j,k}^t\|_2)$. ▷ normalize $\bar{\mathbf{w}}$
      Draw $m \gtrsim W^{5.5}b^{14}\log^5(d/\epsilon)d/(L^4\mu^9\epsilon^{3/2})$ new i.i.d. samples from $\mathcal{D}$
      $\hat{u}_{j,k}^t = \mathrm{argmin}_{u \in \mathcal{U}_{(a,b)}} (1/m)\sum_{i=1}^m (u(\widehat{\mathbf{w}}_{j,k}^t \cdot \mathbf{x}^{(i)}) - y^{(i)})^2$.
      $\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}_{j,k}^t; \hat{u}_{j,k}^t) = (1/m)\sum_{i=1}^m (\hat{u}_{j,k}^t(\widehat{\mathbf{w}}_{j,k}^t \cdot \mathbf{x}^{(i)}) - y^{(i)})\mathbf{x}^{(i)}$.
      $\bar{\mathbf{w}}_{j,k}^{t+1} = \widehat{\mathbf{w}}_{j,k}^t - \eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}_{j,k}^t; \hat{u}_{j,k}^t)$
    end for
    $\mathcal{P}_k \leftarrow \mathcal{P}_k \cup \{(\widehat{\mathbf{w}}_{j,k}^T; \hat{u}_{j,k}^T)\}$.
  end for
  $\mathcal{P} = \cup_{k=1}^{t_0} \mathcal{P}_k \cup \{(\mathbf{w} = 0; u(z) = 0)\}$
end for
$(\widehat{\mathbf{w}}; \hat{u}) = \text{Test}[(\mathbf{w}; u) \in \mathcal{P}]$ (Algorithm 4) ▷ testing
Return: $(\widehat{\mathbf{w}}; \hat{u})$

Algorithm 4 Testing

Input: $\epsilon > 0$; positive parameters: a, b, L, R, W; list of solutions $\mathcal{P}$; let $r \gtrsim \frac{1}{L}\log(bW/(L\epsilon)\log^2(1/\epsilon))$
Draw $m' \gtrsim (bW/L)^4\log^5(1/\epsilon)/\epsilon^2$ new i.i.d. samples from $\mathcal{D}$.
$(\widehat{\mathbf{w}}; \hat{u}) = \mathrm{argmin}_{(\mathbf{w};u)\in\mathcal{P}}\{\frac{1}{m'}\sum_{i=1}^{m'}(u(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})^2\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}^{(i)}| \le Wr\}\}$.
Return: $(\widehat{\mathbf{w}}; \hat{u})$

0 to W to find the (approximate) value of $\|\mathbf{w}^*\|_2$, there must exist an index j^* such that the value of β_{j^*} is $\eta\sqrt{\epsilon}$ close to $\|\mathbf{w}^*\|_2$, i.e., $|\beta_{j^*} - \|\mathbf{w}^*\|_2| \le \eta\sqrt{\epsilon}$. We now consider this $j^{*\mathrm{th}}$ outer loop and ignore the subscript j^* for simplicity. Let $\mathbf{w}^t = \|\mathbf{w}^*\|_2(\bar{\mathbf{w}}^t/\|\bar{\mathbf{w}}^t\|_2)$, which is the true normalized vector of $\bar{\mathbf{w}}^t$ that has no error.

We study the squared distance between $\bar{\mathbf{w}}^{t+1}$ and $\mathbf{w}^*$:

$$\begin{aligned}\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 &= \|\widehat{\mathbf{w}}^t - \eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t) - \mathbf{w}^*\|_2^2 \\ &= \|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2 + \eta^2\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t)\|_2^2 - 2\eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t)\cdot(\widehat{\mathbf{w}}^t - \mathbf{w}^*). \end{aligned} \tag{46}$$

Applying Lemma 4.3 to (46), and plugging in Proposition 3.1, we get that when drawing

$$m \gtrsim \frac{dW^{9/2}b^4\log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}} + \frac{1}{\epsilon\delta}\right), \tag{47}$$

samples from the distribution, it holds with probability at least $1 - \delta$ that:

$$\begin{aligned}\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \le \|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2 &+ \eta^2(10(\mathrm{OPT} + \epsilon) + 4b^2\|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2) \\ &+ 2\eta(2(\mathrm{OPT} + \epsilon)/b + 2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2 - \mu\|\mathbf{v}^t\|_2^2),\end{aligned} \tag{48}$$

where $\mu = Ca^2LR^4/b$ with C being an absolute constant, and where $\mathbf{v}^t$ is the component of $\mathbf{w}^*$ that is orthogonal to $\widehat{\mathbf{w}}^t$, i.e.,

$$\mathbf{v}^t = \mathbf{w}^* - (\mathbf{w}^*\cdot\widehat{\mathbf{w}}^t)\widehat{\mathbf{w}}^t/\|\widehat{\mathbf{w}}^t\|_2^2 = (\mathbf{w}^*)^{\perp_{\widehat{\mathbf{w}}^t}}.$$

Note that $\|\mathbf{v}^t\|_2$ is invariant to the rescaling of $\widehat{\mathbf{w}}^t$, in other words, $\mathbf{w}^*$ has the same orthogonal component $\mathbf{v}^t$ for all $\bar{\mathbf{w}}^t$, $\mathbf{w}^t$ and $\widehat{\mathbf{w}}^t$.

Since $\|\widehat{\mathbf{w}}^t - \mathbf{w}^t\|_2 \le \eta\sqrt{\epsilon}$, we have

$$\|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2 = \|\widehat{\mathbf{w}}^t - \mathbf{w}^t + \mathbf{w}^t - \mathbf{w}^*\|_2^2 \le \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \eta^2\epsilon + 2\eta\sqrt{\epsilon}\|\mathbf{w}^t - \mathbf{w}^*\|_2. \tag{49}$$

In addition, by triangle inequality we have $\|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2 \le \|\mathbf{w}^t - \mathbf{w}^*\|_2 + \eta\sqrt{\epsilon}$. Therefore, substituting $\mathbf{w}^t$ with $\widehat{\mathbf{w}}^t$ in (48), we get:

$$\begin{aligned}
\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 &\le \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \eta^2\epsilon + 2\eta\sqrt{\epsilon}\|\mathbf{w}^t - \mathbf{w}^*\|_2 \\
&\quad + \eta^2(10(\mathrm{OPT}+\epsilon) + 4b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 4b^2\eta^2\epsilon + 8b^2\eta\sqrt{\epsilon}\|\mathbf{w}^t - \mathbf{w}^*\|_2) \\
&\quad + 2\eta(2(\mathrm{OPT}+\epsilon)/b + 2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})(\|\mathbf{w}^t - \mathbf{w}^*\|_2 + \eta\sqrt{\epsilon}) - \mu\|\mathbf{v}^t\|_2^2) \\
&\le \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \eta^2(24(\mathrm{OPT}+\epsilon) + 4b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2) \\
&\quad + 2\eta(2(\mathrm{OPT}+\epsilon)/b + 4(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2 - \mu\|\mathbf{v}^t\|_2^2),
\end{aligned} \tag{50}$$

where we used $4b^2\eta^2 \le 1$, which holds because $\eta = \mu/(4b^2)$.

Our goal is to show that $\|\mathbf{v}^{t+1}\|_2^2 \le \|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \le (1-c)\|\mathbf{v}^t\|_2^2 + \epsilon$, where $c \in (0,1)$ is a constant and ϵ is a small error parameter. However, this linear contraction can only be obtained when $\|\mathbf{v}^t\|_2$ is relatively small compared to $\|\mathbf{w}^*\|_2$. Specifically, as will be manifested in Claim D.3 and the proceeding proof, the linear contraction is achieved only when $\|\mathbf{v}^t\|_2 \le \mu\|\mathbf{w}^*\|_2/(4b)$. Luckily, we can start with a $\mathbf{v}^0$ such that this condition is satisfied, due to the initialization subroutine Algorithm 2, as proved in Lemma 4.1. We prove the following claim.

Claim D.3. Let $\eta = \mu/(4b^2)$. Then, under the assumptions of Theorem 4.2, with probability at least $1-\delta$, we have

$$\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \le \left(1 - \frac{\mu^2}{32b^2}\right)\|\mathbf{v}^t\|_2^2,$$

whenever $\|\mathbf{v}^t\|_2 \ge (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$.

Proof of Claim D.3. Since the norm of $\mathbf{w}^t$ is normalized to $\mathbf{w}^*$, the quantity $\|\mathbf{w}^t - \mathbf{w}^*\|^2$ is controlled by $\|\mathbf{v}^t\|_2^2$. In particular, let $\mathbf{w}^* = \alpha_t\mathbf{w}^t + \mathbf{v}^t$. Then, since $\mathbf{v}^t \perp \mathbf{w}^t$, we have $\|\mathbf{w}^*\|_2^2 = \alpha_t^2\|\mathbf{w}^t\|_2^2 + \|\mathbf{v}^t\|_2^2 = \alpha_t^2\|\mathbf{w}^*\|_2^2 + \|\mathbf{v}^t\|_2^2$, thus, $\alpha_t^2 = 1 - \|\mathbf{v}^t\|_2^2/\|\mathbf{w}^*\|_2^2$, and $\|\mathbf{v}^t\|_2^2 = (1-\alpha_t^2)\|\mathbf{w}^*\|_2^2$. In addition, $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2$ can be expressed as a function of α_t and $\mathbf{w}^*$, as

$$\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 = (1-\alpha_t)^2\|\mathbf{w}^*\|_2^2 + \|\mathbf{v}^t\|_2^2 = 2(1-\alpha_t)\|\mathbf{w}^*\|_2^2. \tag{51}$$

Note that since $\alpha_t = \sqrt{1 - \|\mathbf{v}^t\|_2^2/\|\mathbf{w}^*\|_2^2}$, denoting $\rho_t = \|\mathbf{v}^t\|_2/\|\mathbf{w}^*\|_2$, we further have:

$$1 - \alpha_t = 1 - \sqrt{1 - \|\mathbf{v}^t\|_2^2/\|\mathbf{w}^*\|_2^2} = 1 - \sqrt{1-\rho_t^2} \le \frac{1}{2}\rho_t^2 + \frac{1}{2}\rho_t^4 \le \rho_t^2, \ \forall \rho_t \in [0,1]. \tag{52}$$

Therefore, plugging (51) and (52) back into (50), we get:

$$\begin{aligned}
\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 &\le 2(1-\alpha_t)\|\mathbf{w}^*\|_2^2 + 4b^2\eta^2(2(1-\alpha_t)\|\mathbf{w}^*\|_2^2) + 8\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\sqrt{2(1-\alpha_t)}\|\mathbf{w}^*\|_2 \\
&\quad - 2\eta\mu\|\mathbf{v}^t\|_2^2 + 24\eta^2(\mathrm{OPT}+\epsilon) + 4\eta(\mathrm{OPT}+\epsilon)/b \\
&\le (\rho_t^2 + \rho_t^4)\|\mathbf{w}^*\|_2^2 + 4b^2\eta^2(\rho_t^2 + \rho_t^4)\|\mathbf{w}^*\|_2^2 + 8\sqrt{2}\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\rho_t\|\mathbf{w}^*\|_2 \\
&\quad - 2\eta\mu\|\mathbf{v}^t\|_2^2 + 24\eta^2(\mathrm{OPT}+\epsilon) + 4\eta(\mathrm{OPT}+\epsilon)/b \\
&= (1 + \rho_t^2 + 4b^2\eta^2(1+\rho_t^2))\|\mathbf{v}^t\|_2^2 + 12\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{v}^t\|_2 - 2\eta\mu\|\mathbf{v}^t\|_2^2 \\
&\quad + 4(6\eta^2 + \eta/b)(\mathrm{OPT}+\epsilon) \\
&\le (1 + \rho_t^2 + 4b^2\eta^2(1+\rho_t^2))\|\mathbf{v}^t\|_2^2 + 12\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{v}^t\|_2 - 2\eta\mu\|\mathbf{v}^t\|_2^2 + 5\eta(\mathrm{OPT}+\epsilon),
\end{aligned} \tag{53}$$

where in the last inequality we observed that since $\eta = \frac{\mu}{4b^2}$, it holds that $24\eta \le 1$, as μ is small and $b \ge 1$.

Note that we have assumed that $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, which indicates

$$12\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{v}^t\|_2 \leq \frac{1}{8}\eta\mu\|\mathbf{v}^t\|_2^2,$$

since $b \geq 1$, assuming without loss of generality. Furthermore, when $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, it also holds that

$$\frac{1}{8}\eta\mu\|\mathbf{v}^t\|_2^2 \geq \frac{(96)^2}{8\mu^2}\eta\mu(\mathrm{OPT} + \epsilon) \geq 5\eta(\mathrm{OPT} + \epsilon),$$

since we have assumed $\mu = Ca^2LR^4/b \leq 1$ without loss of generality. Finally, as we will show in the rest of the proof, it holds that $\|\mathbf{v}^{t+1}\|_2 \leq \|\mathbf{v}^t\|_2$ for $t = 0, 1, \ldots, T$, thus as $\eta = \mu/(4b^2)$, we have $\|\mathbf{v}^t\|_2 \leq \sqrt{\eta\mu}\|\mathbf{w}^*\|_2/2 = \mu\|\mathbf{w}^*\|_2/(4b)$, since $\|\mathbf{v}^0\|_2 \leq \sqrt{\eta\mu}\|\mathbf{w}^*\|_2/2$. This condition guarantees that

$$\rho_t^2 = \|\mathbf{v}^t\|_2^2/\|\mathbf{w}^*\|_2^2 \leq \frac{1}{4}\eta\mu.$$

Plugging these conditions back into (53), it is then simplified as (note that $1 + \rho_t^2 \leq 1 + (1/4)\eta\mu \leq 9/8$ for $\eta\mu \leq 1/2$):

$$\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \leq \left(1 + \frac{9}{2}b^2\eta^2 - \frac{3}{2}\eta\mu\right)\|\mathbf{v}^t\|_2^2.$$

Therefore, when $\eta = \mu/(4b^2)$ we have

$$\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \leq \left(1 - \frac{\mu^2}{32b^2}\right)\|\mathbf{v}^t\|_2^2,$$

completing the proof. □

We proceed first under the condition that $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$ holds for $t = 0, \ldots, T$ and show that after some certain number of iterations T this condition must be violated. Observe if $\|\mathbf{v}^t\|_2 \leq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, then it holds $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \lesssim (1/\mu^2)(\mathrm{OPT} + \epsilon)$, implying that $\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x})$ is a hypothesis achieving constant approximation error according to Claim 4.4, hence the algorithm can be terminated. However, note that T only works as an upper bound for the iteration complexity of or algorithm, and it is possible that the condition $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$ is violated at some step $t^* < T$. However, we will show later that the value of $\|\mathbf{v}^T\|_2$ can not be larger than $c\|\mathbf{v}^{t^*}\|_2$ where c is an absolute constant. We observe that:

$$\mathbf{v}^{t+1} = \mathbf{w}^* - (\mathbf{w}^* \cdot \mathbf{w}^{t+1})\mathbf{w}^{t+1}/\|\mathbf{w}^{t+1}\|_2^2 = \mathbf{w}^* - (\mathbf{w}^* \cdot \bar{\mathbf{w}}^{t+1})\bar{\mathbf{w}}^{t+1}/\|\bar{\mathbf{w}}^{t+1}\|_2^2 = (\mathbf{w}^*)^{\perp_{\bar{\mathbf{w}}^{t+1}}},$$

therefore, $\|\mathbf{v}^{t+1}\|_2^2 \leq \|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2$, which, combined with Claim D.3, yields

$$\|\mathbf{v}^{t+1}\|_2^2 \leq \left(1 - \frac{\mu^2}{32b^2}\right)\|\mathbf{v}^t\|_2^2 \leq \left(1 - \frac{\mu^2}{32b^2}\right)^t\|\mathbf{v}^0\|_2^2 \leq \exp\left(-\frac{\mu^2 t}{32b^2}\right)2W^2.$$

The above contraction only holds when $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$. Hence, after at most

$$T = O\left(\frac{b^2}{\mu^2}\log\left(\frac{\mu W}{\epsilon}\right)\right)$$

inner iterations, the algorithm outputs a vector $\mathbf{w}^{t^*}$ with $\|\mathbf{v}^{t^*}\|_2 \leq \frac{96}{\mu}(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, where $t^* \in [T]$.

Now suppose at step $t^* < T$ it holds that $\|\mathbf{v}^{t^*}\|_2 \leq 96(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})/\mu$ but at the next iteration $\|\mathbf{v}^{t^*+1}\|_2 \geq 96(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})/\mu$. Recall first that in Lemma 4.3 we showed that $\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t)\|_2^2 \leq 4b^2\|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2 + 10(\mathrm{OPT} + \epsilon)$. Therefore, revisiting the updating scheme of the algorithm we have

$$\begin{aligned}
\|\mathbf{v}^{t^*+1}\|_2^2 \leq \|\bar{\mathbf{w}}^{t^*+1} - \mathbf{w}^*\|_2^2 &= \|\widehat{\mathbf{w}}^{t^*} - \eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^{t^*}; \hat{u}^{t^*}) - \mathbf{w}^*\|_2^2 \\
&\leq 2\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^*\|_2^2 + 2\eta^2\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^{t^*}; \hat{u}^{t^*})\|_2^2 \\
&\leq (2 + 8b^2\eta^2)\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^*\|_2^2 + 20\eta^2(\mathrm{OPT} + \epsilon) \\
&\leq 3\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^*\|_2^2 + (\mathrm{OPT} + \epsilon),
\end{aligned}$$

where in the last inequality we plugged in the value of $\eta = \mu/(4b^2)$, and used the assumption that $\mu \leq 1$ and $b \geq 1$, hence $20\eta^2 \leq 1$ and $8b^2\eta^2 \leq 1$. Furthermore, recall that by the construction of the grid, $\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^t\|_2 \leq \eta\sqrt{\epsilon}$, implying that $\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^*\|_2^2 \leq 2\|\mathbf{w}^{t^*} - \mathbf{w}^*\|_2^2 + 2\eta^2\epsilon$ by triangle inequality. Therefore, going back to the inequality of $\|\mathbf{v}^{t^*+1}\|_2^2$ above, we get

$$\|\mathbf{v}^{t^*+1}\|_2^2 \leq 6\|\mathbf{w}^{t^*} - \mathbf{w}^*\|_2^2 + 6\eta^2\epsilon + \mathrm{OPT} + \epsilon \leq 6\|\mathbf{w}^{t^*} - \mathbf{w}^*\|_2^2 + 2(\mathrm{OPT} + \epsilon).$$

Finally, observe that since $\|\mathbf{w}^{t^*}\|_2 = \|\mathbf{w}^*\|_2$, it holds $\|\mathbf{w}^{t^*} - \mathbf{w}^*\|_2 \leq \sqrt{2}\|\mathbf{v}^{t^*}\|_2$, hence, we get

$$\|\mathbf{v}^{t^*+1}\|_2^2 \leq 12\|\mathbf{v}^{t^*}\|_2^2 + 2(\mathrm{OPT} + \epsilon).$$

Now since $\|\mathbf{v}^{t^*+1}\|_2 \geq 96(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})/\mu$, the value of $\|\mathbf{v}^t\|_2^2$ will start to decrease again for $t \geq t^* + 1$. This implies that the value of $\|\mathbf{v}^T\|_2$ satisfies

$$\|\mathbf{v}^T\|_2 \leq \sqrt{12}\|\mathbf{v}^{t^*}\|_2 + \sqrt{2}(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon}) \leq \frac{384}{\mu}(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon}).$$

Combining Claim 4.4 and Lemma F.4, as we have guaranteed that $\|\mathbf{v}^T\|_2 \leq (384/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, the hypothesis $\hat{u}^T(\widehat{\mathbf{w}}^T \cdot \mathbf{x})$ has the L_2^2 error that can be bounded as:

$$\mathcal{L}_2(\widehat{\mathbf{w}}^T; \hat{u}^T) \leq 6\mathrm{OPT} + 3b^2(4\|\mathbf{v}^T\|_2^2 + \eta^2\epsilon) + \epsilon = O\left(\frac{b^2}{\mu^2}(\mathrm{OPT} + \epsilon)\right).$$

Setting $\epsilon \leftarrow C'(b^2/\mu^2)\epsilon$ for some universal absolute constant C' we get $\mathcal{L}_2(\widehat{\mathbf{w}}^T; \hat{u}^T) \leq O((b^2/\mu^2)\mathrm{OPT}) + \epsilon$.

It still remains to determine the batch size as drawing a sample set of size m as displayed in (47) only guarantees that the contraction of $\|\mathbf{v}^t\|_2$ at step t holds with probability $1 - \delta$. Applying a union bound on all $t_0 JT = O(\frac{Wb^{10}}{\mu^9\sqrt{\epsilon}}\log(1/\epsilon))$ iterations yields that the contraction holds at every step with probability at least $1 - t_0 JT\delta$. Therefore, setting $\delta \leftarrow \delta(t_0 JT)$ and bringing the value of δ back to (47), we get that it suffices to choose the batch size as:

$$m = \Theta\left(\frac{W^{9/2}b^4\log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}} + \frac{Wb^{10}}{\mu^9\epsilon^{3/2}\delta}\right)\right) = \Theta\left(\frac{W^{5.5}b^{14}d\log^5(d/(\epsilon\delta))}{L^4\mu^9\delta\epsilon^{3/2}}\right),$$

to guarantee that we get a $O(\mathrm{OPT}) + \epsilon$-solution with probability at least $1 - \delta$.

The argument above justifies the claim that among all $t_0 J = Wb^6\log(b/\mu)/(\eta\mu^6\sqrt{\epsilon})$ hypotheses in $\mathcal{P} = \{(\widehat{\mathbf{w}}_j^T; \hat{u}_j^T)\}_{j=1}^{t_0 J}$, there exists at least one hypothesis that achieves L_2^2 error $O(\mathrm{OPT}) + \epsilon$. To select the correct hypothesis from the set $\mathcal{P}$, one only needs to draw a new batch of $m' = \tilde{\Theta}(b^4W^4\log(1/\delta)/(L^4\epsilon^2))$ i.i.d. samples from $\mathcal{D}$, and choose the hypothesis from $\mathcal{P}$ that achieves the minimal empirical error defined in Algorithm 3. To be concrete, we prove the following claim, whose proof can be found in Appendix D.5.

Claim D.4. Fix some positive real numbers μ, ϵ, δ. Let $r = \frac{1}{L}\log(\frac{Cb^4W^4}{L^6\epsilon^2}\log^2(\frac{bW}{\epsilon}))$ where C is a large absolute constant. Given a set of parameter-activation pairs $\mathcal{P} = \{(\mathbf{w}_j; u_j)\}_{j=1}^{t_0 J}$ such that $\mathbf{w}_j \in \mathbb{B}(W)$ and $u_j \in \mathcal{U}_{(a,b)}$ for $j \in [t_0 J]$, where $t_0 J = 4b^9W/(\mu^8\sqrt{\epsilon})$, we have that using

$$m' = \Theta\left(\frac{b^4W^4\log(1/\delta)}{L^4\epsilon^2}\log^5\left(\frac{bW}{L\mu\epsilon}\right)\right),$$

i.i.d. samples from $\mathcal{D}$, for any $(\mathbf{w}_j; u_j) \in \mathcal{P}$ it holds with probability at least $1 - \delta$,

$$\left|\frac{1}{m'}\sum_{i=1}^{m'}(u_j(\mathbf{w}_j \cdot \mathbf{x}^{(i)}) - y^{(i)})^2\mathbb{1}\{|\mathbf{w}_j \cdot \mathbf{x}^{(i)}| \leq Wr\} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j \cdot \mathbf{x}) - y)^2]\right| \leq 2\epsilon.$$

Therefore, according to Claim D.4, we know that testing each $(\widehat{\mathbf{w}}_j^T; \hat{u}_j^T) \in \mathcal{P}$ on a fresh set of m' samples and choosing the one that achieves minimum error yields a solution $(\widehat{\mathbf{w}}; \hat{u})$ that introduces at most 2ϵ error with high probability. In

conclusion, it holds by a union bound that the Algorithm 3 delivers a solution with $O(\mathrm{OPT}) + \epsilon$ error with probability at least $1 - 2\delta$. The total sample complexity of our algorithm is

$$N = t_0 JTm + m' = \Theta\left(\frac{W^{6.5} b^{24} d \log^5(d/(\epsilon\delta))}{L^4 \mu^{18} \delta \epsilon^2} + \frac{b^4 W^4 \log(1/\delta) \log^5(1/\epsilon)}{L^4 \epsilon^2}\right) = \Theta\left(\frac{W^{6.5} b^{24} d \log^6(d/(\epsilon\delta))}{L^4 \mu^{18} \delta \epsilon^2}\right).$$

Choosing $\delta = 1/6$ above, we get that the Algorithm 3 succeeds with probability at least $1 - 2\delta = 2/3$, completing the proof of Theorem D.2. □

D.3. Proof of Lemma 4.3

This subsection is devoted to the proof of Lemma 4.3. To this aim, we first show the following lemmas that bound from above the norm of the population gradient $\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)$ and the difference between the population gradient and the empirical gradient $\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)$.

Lemma D.5. *Let S be a sample set of m i.i.d. samples of size at least $m \gtrsim d \log^4(d/(\epsilon\delta))(b^2 W^3/L^2\epsilon)^{3/2}$. Furthermore, given $\mathbf{w}^t \in \mathbb{B}(W)$, let $\hat{u}^t$ be defined as in (P). Then, it holds that with probability at least $1 - \delta$,*

$$\|\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2^2 \leq 8(\mathrm{OPT} + \epsilon) + 2b^2 \|\mathbf{w}^t - \mathbf{w}^*\|_2^2.$$

Proof. By the definition of ℓ_2 norms, we have:

$$\begin{aligned}
\|\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 &= \max_{\|\mathbf{v}\|_2 = 1} \nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot \mathbf{v} \\
&= \max_{\|\mathbf{v}\|_2 = 1} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)\mathbf{v} \cdot \mathbf{x}] \\
&= \max_{\|\mathbf{v}\|_2 = 1} \Bigg\{ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}) + u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))(\mathbf{v} \cdot \mathbf{x})] \\
&\quad + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}) + u^*(\mathbf{w}^* \cdot \mathbf{x}) - y)(\mathbf{v} \cdot \mathbf{x})] \Bigg\}.
\end{aligned}$$

By the Cauchy-Schwarz inequality, we further have:

$$\begin{aligned}
&\|\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 \\
&\leq \max_{\|\mathbf{v}\|_2 = 1} \Bigg\{ \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2] \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{v} \cdot \mathbf{x})^2]} + \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{v} \cdot \mathbf{x})^2]} \\
&\quad + \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{v} \cdot \mathbf{x})^2]} + \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^*(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{v} \cdot \mathbf{x})^2]} \Bigg\} \\
&\leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2]} + \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2]} \\
&\quad + \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]} + \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^*(\mathbf{w}^* \cdot \mathbf{x}) - y)^2]},
\end{aligned}$$

where in the last inequality we used the fact that $\mathcal{D}_\mathbf{x}$ is isotropic hence $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{v} \cdot \mathbf{x})^2] = 1$. It remains to bound these four terms above respectively. The first term is bounded by $\sqrt{\epsilon}$ for every $\mathbf{w}^t \in \mathbb{B}(W)$, with probability at least $1 - \delta$ according to Lemma F.4. The fourth term is bounded by OPT, by definition. Recall that in Lemma 3.3 we showed that the second term in the display above is upper-bounded: $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \leq \mathrm{OPT}$. For the third term, note that $u^{*t} \in \mathrm{argmin}_{u \in \mathcal{U}_{(a,b)}} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]$, therefore, since $u^* \in \mathcal{U}_{(a,b)}$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^*(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq b^2 \|\mathbf{w}^t - \mathbf{w}^*\|_2^2,$$

after applying the fact that u^* is b-Lipschitz. Thus, in conclusion, we have

$$\|\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 \leq 2\sqrt{\mathrm{OPT}} + \sqrt{\epsilon} + b\|\mathbf{w}^t - \mathbf{w}^*\|_2.$$

Furthermore, since $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, we get with probability at least $1 - \delta$:

$$\|\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2^2 \leq 8\mathrm{OPT} + 8\epsilon + 2b^2 \|\mathbf{w}^t - \mathbf{w}^*\|_2^2,$$

completing the proof of Lemma D.5. □

Now we prove that the distance between $\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)$ and $\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)$ is bounded by $b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+\mathrm{OPT}+\epsilon$ with high probability.

Lemma D.6. *Let S be a sample set of $m \gtrsim (dW^{9/2}b^4\log^4(d/(\epsilon\delta))/L^4)(1/\epsilon^{3/2}+1/(\epsilon\delta))$ i.i.d. samples. Given a vector $\mathbf{w}^t \in \mathbb{B}(W)$, it holds that with probability at least $1-\delta$,*

$$\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)-\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)\|_2 \le \sqrt{b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+\mathrm{OPT}+\epsilon}.$$

Proof. Since for any mean-zero independent random variables $\mathbf{z}_j$, we have $\mathbf{E}[\|\sum_j \mathbf{z}_j\|_2^2] = \sum_j \mathbf{E}[\|\mathbf{z}_j\|_2^2]$, hence, by Markov:

$$\mathbf{Pr}[\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)-\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)\|_2 \ge s] \le \frac{1}{ms^2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)\mathbf{x}\|_2^2]\,. \tag{54}$$

By linearity of expectation, we have:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)\mathbf{x}\|_2^2] = \sum_{k=1}^{d}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2(\mathbf{x}_k)^2],$$

where $\mathbf{x}_k = \mathbf{e}_k\cdot\mathbf{x}$ and $\mathbf{e}_k$ is the k^{th} unit basis of $\mathbb{R}^d$. Let $r = O(W/L\log(1/(L\epsilon')))$, then it holds $\mathbf{Pr}[|\mathbf{x}_k|\ge r]\le\epsilon'$. Then, the variance above can be decomposed into the following parts:

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2] &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\ge r\}]\\ &\quad+\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\le r\}].\end{aligned}$$

Since $|y|\le M = O(bW/L\log(bW/\epsilon))$, and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}^t\cdot\mathbf{x})^4\mathbf{x}_k^4]\le W^4c^2/L^8$, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x}_k^4]\le c^2/L^4$ for $\mathcal{D}_\mathbf{x}$ is L-sub-exponential, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\ge r\}] &\le 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}))^2+y^2)\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\ge r\}]\\ &\le 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(b(\mathbf{w}^t\cdot\mathbf{x}))^2+y^2)\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\ge r\}]\\ &\le 2b^2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t\cdot\mathbf{x})^4\mathbf{x}_k^4]\,\mathbf{Pr}[|\mathbf{x}_k|\ge r]}\\ &\quad+2M^2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x}_k^4]\,\mathbf{Pr}[|\mathbf{x}_k|\ge r]}\\ &\le (2cb^2W^2/L^4)\sqrt{\epsilon'}+(2cM^2/L^2)\sqrt{\epsilon'}\le(4cM^2/L^2)\sqrt{\epsilon'}.\end{aligned} \tag{55}$$

In addition, $(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2$ can be decomposed as the following:

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2] &\le 4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2]+4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]\\ &\quad+4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]+4\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^*(\mathbf{w}^*\cdot\mathbf{x})-y)^2].\end{aligned}$$

The first term is upper bounded by 4ϵ with probability at least $1-\delta$ for every $\mathbf{w}^t\in\mathbb{B}(W)$ whenever $m\gtrsim d\log^4(d/(\epsilon\delta))(b^2W^3/L^2\epsilon)^{3/2}$, as proved in Lemma F.4. The second term is smaller than $4\mathrm{OPT}$, which is shown in Lemma 3.3. The third term can be upper bounded using again the definition of $u^{*t} = \mathrm{argmin}_{u\in\mathcal{U}_{(a,b)}}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]$, as

$$4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\le 4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^*(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\le 4b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2,$$

using the fact that u^* is b-Lipschitz and $\mathcal{D}_\mathbf{x}$ is isotropic. Lastly, the fourth term is bounded by $4\mathrm{OPT}$ by the definition of $u^*(\mathbf{w}^*\cdot\mathbf{x})$. In summary, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\le r\}] &\le r^2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2]\\ &\le 4r^2(b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+2\mathrm{OPT}+\epsilon),\end{aligned}$$

which, combining with (55), implies that the expectation $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2]$ is bounded by:

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2] &\le 4r^2b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+4r^2(2\mathrm{OPT}+2\epsilon)\\ &\le \frac{CW^2}{L^2}\log^2\left(\frac{b}{L\epsilon}\right)(b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+\mathrm{OPT}+\epsilon),\end{aligned}$$

where C is a large absolute constant. Note to get the inequality above we chose $\epsilon'=C\epsilon^2(L/b)^4$, which then indicates that $4c(M/L)^2\sqrt{\epsilon'}\le r^2\epsilon$. Summing the inequality above from $k=1$ to d delivers the final upper bound on the variance:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)\mathbf{x}\|_2^2]\le\frac{dCW^2}{L^2}\log^2\left(\frac{b}{L\epsilon}\right)(b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+\mathrm{OPT}+\epsilon).$$

Thus, plugging the upper bound on the variance above back to (54), as long as $m\gtrsim(dW^2/L^2)\log^2(b/(L\epsilon))/\delta$, we get with probability at least $1-\delta$,

$$\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)-\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)\|_2\le\sqrt{b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+\mathrm{OPT}+\epsilon}.$$

Noting that $m\gtrsim(dW^{9/2}b^4\log^4(d/(\epsilon\delta))/L^4)(1/\epsilon^{3/2}+1/(\epsilon\delta))$ certainly satisfies the condition on m above as $m\gtrsim(dW^2/L^2)\log^2(b/(L\epsilon))/\delta$, thus, we completed the proof of Lemma D.6 □

We can now proceed to the proof of Lemma 4.3, which can be derived directly from the preceding lemmas.

Lemma D.7 (Upper Bound on Empirical Gradient Norm)**.** *Let S be a set of i.i.d. samples of size $m\gtrsim(dW^{9/2}b^4\log^4(d/(\epsilon\delta))/L^4)(1/\epsilon^{3/2}+1/(\epsilon\delta))$. Given any $\mathbf{w}^t\in\mathbb{B}(W)$, let $\hat{u}^t\in\mathcal{U}_{(a,b)}$ be the solution of optimization problem (P) with respect to $\mathbf{w}^t$ and sample set S. Then, with probability at least $1-\delta$, we have that $\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)\|_2^2\le 4b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+10(\mathrm{OPT}+\epsilon)$.*

Proof. The lemma follows directly by combining Lemma D.5, Lemma D.6 and the triangle inequality. □

D.4. Proof of Claim 4.4

We restate and prove Claim 4.4.

Claim D.8. Let $\mathbf{w}$ be any vector from $\mathbb{B}$. Let $\hat{u}_{\mathbf{w}}$ be the activation defined as the empirical-optimal solution of the optimization problem (P) for a fixed parameter vector $\mathbf{w}\in\mathbb{R}^d$ with batch size $m\gtrsim d\log^4(d/(\epsilon\delta))(b^2W^3/(L^2\epsilon))^{3/2}$. Then the L_2^2 error of $\hat{u}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})$ is bounded by: $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-y)^2]\le 8(\mathrm{OPT}+\epsilon)+4b^2\|\mathbf{w}-\mathbf{w}^*\|_2^2$.

Proof. Let $u^*_{\mathbf{w}}$, $u_{\mathbf{w}}$ be the optimal activation of problem (EP*) and (EP) under parameter $\mathbf{w}$ respectively. Then, direct calculation gives:

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}_{\mathbf{w}}-y)^2]\\ &=\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})+u_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})+u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x})+u^*(\mathbf{w}^*\cdot\mathbf{x})-y)^2]\\ &\le 4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\hat{u}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x}))^2]+4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(u_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x}))^2]\\ &\quad+4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]+4\mathrm{OPT}\\ &\le 8(\mathrm{OPT}+\epsilon)+4b^2\|\mathbf{w}-\mathbf{w}^*\|_2^2,\end{aligned}\tag{56}$$

where in the second inequality we used the results from Lemma 3.3, Lemma F.4 and we applied the observation that:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\le\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(u^*(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\le b^2\|\mathbf{w}-\mathbf{w}^*\|_2^2,$$

by the definition of $u^*_{\mathbf{w}}$. □

D.5. Proof of Claim D.4

We restate Claim D.4 and show the number of samples needed for the testing subroutine Algorithm 4.

Claim D.9. Fix some positive real numbers μ, ϵ, δ. Let $r = \frac{1}{L}\log(\frac{Cb^4W^4}{L^6\epsilon^2}\log^2(\frac{bW}{\epsilon}))$ where C is a large absolute constant. Given a set of parameter-activation pairs $\mathcal{P} = \{(\mathbf{w}_j; u_j)\}_{j=1}^{t_0 J}$ such that $\mathbf{w}_j \in \mathbb{B}(W)$ and $u_j \in \mathcal{U}_{(a,b)}$ for $j \in [t_0 J]$, where $t_0 J = 4b^9 W/(\mu^8\sqrt{\epsilon})$, we have that using

$$m' = \Theta\bigg(\frac{b^4W^4\log(1/\delta)}{L^4\epsilon^2}\log^5\left(\frac{bW}{L\mu\epsilon}\right)\bigg),$$

i.i.d. samples from $\mathcal{D}$, for any $(\mathbf{w}_j; u_j) \in \mathcal{P}$ it holds with probability at least $1-\delta$,

$$\left|\frac{1}{m'}\sum_{i=1}^{m'}(u_j(\mathbf{w}_j\cdot\mathbf{x}^{(i)}) - y^{(i)})^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}^{(i)}| \le Wr\} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x}) - y)^2]\right| \le 2\epsilon.$$

Proof. Fix some $r \ge 0$, and fix any $(\mathbf{w}_j, u_j) \in \mathcal{P}$. Since $\mathcal{D}_\mathbf{x}$ is sub-exponential, we have $\mathbf{Pr}[|\mathbf{w}_j\cdot\mathbf{x}| \ge \|\mathbf{w}_j\|_2 r] \le \frac{1}{L^2}\exp(-Lr)$. Consider random variables $Z_{i,j} = (u_j(\mathbf{w}_j\cdot\mathbf{x}^{(i)}) - y^{(i)})^2\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}^{(i)}| \le r\}$, $i = 1, \cdots, m$, $j = 1, \cdots, t_0 J$, where $(\mathbf{x}^{(i)}, y^{(i)})$ are independent random variables drawn from $\mathcal{D}$. Recall that using the result from Fact F.3, we can truncate the labels y such that $|y| \le M$, where $M = C(bW/L)\log(bW/\epsilon)$ for some large absolute constant C. Hence, $|Z_{i,j}| \le 2(u_j^2(\mathbf{w}_j\cdot\mathbf{x}^{(i)}) + (y^{(i)})^2)\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}^{(i)}| \le Wr\} \le 2(b^2W^2r^2 + M^2)$, note that we used the assumption that u is b-Lipschitz in the last inequality. Therefore, applying Hoeffding's inequality on $Z_{i,j}$ we get:

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{m'}(Z_{i,j} - E[Z_{i,j}])\right| \ge m't\right] \le \exp\bigg(-\frac{cm't^2}{(b^2W^2r^2 + M^2)^2}\bigg),$$

where c is an absolute constant. Since there are $t_0 J = Wb^6/(\mu^6\eta\sqrt{\epsilon}) = 4b^8W/(\mu^7\sqrt{\epsilon})$ elements in the set $\mathcal{P}$, thus applying a union bound yields:

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{m'} Z_{i,j} - E[Z_{i,j}]\right| \ge m't, \forall j \in [J]\right] \le \exp\bigg(-\frac{cm't^2}{(b^2W^2r^2 + M^2)^2} + \log(4b^8W/(\mu^7\sqrt{\epsilon}))\bigg).$$

Therefore, when

$$m' = \frac{(b^2W^2r^2 + M^2)^2}{c\epsilon^2}\left(\log\left(\frac{4b^8W}{\mu^7\sqrt{\epsilon}}\right) + \log(1/\delta)\right), \tag{57}$$

we have that with probability at least $1-\delta$:

$$\left|\frac{1}{m'}\sum_{i=1}^{m'}(u_j(\mathbf{w}_j\cdot\mathbf{x}^{(i)}) - y^{(i)})^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}^{(i)}| \le Wr\} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x}) - y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}| \le Wr\}]\right| \le \epsilon, \tag{58}$$

for any $(\mathbf{w}_j, u_j) \in \mathcal{P}$. In addition, as $\mathbf{Pr}[|\mathbf{w}_j\cdot\mathbf{x}| \ge Wr] \le \mathbf{Pr}[|\mathbf{w}_j\cdot\mathbf{x}| \ge \|\mathbf{w}_j\|_2 r] \le \frac{2}{L^2}\exp(-Lr)$, let $\epsilon' = \frac{2}{L^2}\exp(-Lr)$, we further have:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x}) - y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}| \ge Wr\}] \\
&\le 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((u_j(\mathbf{w}_j\cdot\mathbf{x}))^2 + M^2)\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}| \ge Wr\}] \\
&\le 2b^2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}_j\cdot\mathbf{x})^4]\,\mathbf{Pr}[|\mathbf{w}_j\cdot\mathbf{x}| \ge Wr]} + M^2\,\mathbf{Pr}[|\mathbf{w}_j\cdot\mathbf{x}| \ge Wr] \\
&\le 2cb^2(W/L)^2\sqrt{\epsilon'} + M^2\epsilon' \le (2cb^2(W/L)^2 + M^2)\sqrt{\epsilon'},
\end{aligned}$$

where in the second inequality we use Cauchy-Schwarz inequality and in the last inequality we used the property that for any unit vector $\mathbf{a}$ it holds $\mathbf{E}[(\mathbf{a}\cdot\mathbf{x})^4] \le c^2/L^4$ for some absolute constant c as $\mathbf{x}$ possesses a $\frac{1}{L}$-sub-exponential

tail. Therefore, choosing $r = \frac{1}{L}\log(\frac{C^2b^4W^4}{L^6\epsilon^2}\log^2(\frac{bW}{\epsilon})) = \tilde{O}(\frac{1}{L}\log(\frac{bW}{L\epsilon}))$ for some large absolute constant C renders $\sqrt{\epsilon'} \leq \epsilon/(2Cb^2(W/L)^2\log^2(bW/\epsilon))$, and we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr\}]\leq\epsilon.$$

Observe that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2]$ is the sum of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr\}]$ and $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\leq Wr\}]$, we thus have

$$\begin{aligned}0&\leq\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2]-\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\leq Wr\}]\\&\leq\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr\}]\leq\epsilon.\end{aligned}$$

Bringing the choice of r back to (57) we get that it is sufficient to choose m' as

$$m'=\frac{C\log(\log(1/\epsilon))}{\epsilon^2}\left(b^2\left(\frac{W}{L}\right)^2\log^2\left(\frac{bW}{L\epsilon^2}\right)\right)^2\left(\log\left(\frac{4b^8W}{\mu^7\sqrt{\epsilon}}\right)+\log(1/\delta)\right)=\tilde{\Theta}\left(\frac{b^4W^4\log(1/\delta)}{L^4\epsilon^2}\log^5\left(\frac{bW}{L\mu\epsilon}\right)\right).$$

Therefore, using $m'=\tilde{\Omega}(b^4W^4/(L^4\epsilon^2))$ samples, (58) indicates that with probability at least $1-\delta$, for any $(\mathbf{w}_j,u_j)\in\mathcal{P}$ it holds

$$\begin{aligned}&\left|\frac{1}{m'}\sum_{i=1}^{m'}(u_j(\mathbf{w}_j\cdot\mathbf{x}^{(i)})-y^{(i)})^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}^{(i)}|\leq Wr\}-\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2]\right|\\&\leq\left|\frac{1}{m'}\sum_{i=1}^{m'}(u_j(\mathbf{w}_j\cdot\mathbf{x}^{(i)})-y^{(i)})^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}^{(i)}|\leq Wr\}-\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\leq Wr\}]\right|\\&\quad+\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2]-\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\leq Wr\}]\right|\\&\leq 2\epsilon,\end{aligned}$$

thus completing the proof of Claim D.4. □

E. Efficiently Computing the Optimal Empirical Activation

In this section, we show that the optimization problem (P) can be solved efficiently, following the framework from (Lu & Hochbaum, 2022) with minor modifications. We show that, for any $\epsilon>0$, there is an efficient algorithm that runs in $\tilde{O}(m^2\log(1/\epsilon))$ time and outputs a solution $\hat{v}^t(z)$ such that $\|\hat{v}^t(z)-\hat{u}^t(z)\|_\infty\leq\epsilon$.

Proposition E.1 (Approximating the Optimal Empirical Activation)**.** *Let $\epsilon>0$, and $\mathcal{D}_\mathbf{x}$ be (L,R)-well behaved. Let $\hat{u}^t\in\mathcal{U}_{(a,b)}$ be the optimal solution of the optimization problem Equation (P) given a sample set S of size m drawn from $\mathcal{D}$ and a parameter $\mathbf{w}^t\in\mathbb{B}(W)$. There exists an algorithm that produces an activation $\hat{v}^t\in\mathcal{U}_{(a,b)}$ such that $\|\hat{v}^t-\hat{u}^t\|_\infty\leq\epsilon$, with computation complexity $\tilde{O}(m^2\log(bW/(L\epsilon)))$.*

To show Proposition E.1, we leverage the following result:

Lemma E.2 (Section 5 (Lu & Hochbaum, 2022))**.** *Let $f_i(y)$ and $h_i(y)$ be any convex lower semi-continuous functions for $i=1,\ldots,m$. Consider the following convex optimization problem*

$$(\hat{y}_1,\ldots,\hat{y}_m)=\operatorname*{argmin}_{y_1,\ldots,y_m}\sum_{i=1}^{m}f_i(y_i)+\sum_{i=1}^{m-1}h_i(y_i-y_{i+1}),\tag{59}$$

where $y_i\in[-U,U]$ for all $i=1,\ldots,m$ for some positive constant U. Then, for any $\epsilon>0$, there exists an algorithm (the cc-algorithm (Lu & Hochbaum, 2022)) that outputs an ϵ-close solution $\{y_1,\ldots,y_m\}$ such that $|y_i-\hat{y}_i|\leq\epsilon$ for all $i\in[m]$ with computational complexity $O(m^2\log(U/\epsilon))$.

Proof of Proposition E.1. We first formulate problem (P) into a quadratic optimization problem with linear constraints. To guarantee that $\hat{u}^t$ is an element in $\mathcal{U}_{(a,b)}$ that satisfies $\hat{u}^t(0) = 0$, we add a zero point $(\mathbf{x}^{(0)}, y^{(0)}) = (\mathbf{0}, 0)$ to the data set S if S does not contain $(\mathbf{0}, 0)$ in the first place. We will thus assume without loss of generality that the data set contains $(\mathbf{0}, 0)$. Denote $z_i = \mathbf{w} \cdot \mathbf{x}^{(i)}$ such that $z_1 \leq z_2 \leq \cdots \leq z_m$ after rearranging the order of $(\mathbf{x}^{(i)}, y^{(i)})$'s, and suppose $z_k = \mathbf{w} \cdot \mathbf{x}_0 = 0$ for a $k \in [m]$. Then (P) is equivalent to the following optimization problem:

$$
\begin{aligned}
(\hat{y}^{(1)}, \cdots, \hat{y}^{(m)}) = \underset{\tilde{y}^{(i)}, i\in[m]}{\operatorname{argmin}} \; & \sum_{i=1}^{m} (\tilde{y}^{(i)} - y^{(i)})^2 \\
\text{s.t. } & 0 \leq \tilde{y}^{(i+1)} - \tilde{y}^{(i)}, \; 1 \leq i \leq k-1, \\
& a(z_{i+1} - z_i) \leq \tilde{y}^{(i+1)} - \tilde{y}^{(i)}, \; 1 \leq i \leq k-1, \\
& \tilde{y}^{(i+1)} - \tilde{y}^{(i)} \leq b(z_{i+1} - z_i), \; 1 \leq i \leq m-1, \\
& \tilde{y}^{(k)} = 0.
\end{aligned}
$$

Define $h_i(y) = \mathcal{I}_{[-b(z_{i+1}-z_i), 0]}(y)$ for $i = 1 \ldots, k-1$, $h_i(y) = \mathcal{I}_{[-b(z_{i+1}-z_i), -a(z_{i+1}-z_i)]}(y)$ for $i = k \ldots, m-1$, where $\mathcal{I}_{\mathcal{Y}}(y)$ is the indicator function of a convex set $\mathcal{Y}$, i.e., $\mathcal{I}_{\mathcal{Y}}(y) = 0$ if $y \in \mathcal{Y}$ and $\mathcal{I}_{\mathcal{Y}}(y) = +\infty$ otherwise. It is known that h_i's are convex and sub-differentiable on their domain $\mathcal{Y}_i$. In addition, let $f_i(y) = \frac{1}{2}(y - y^{(i)})^2$ for $i \neq k$ and $f_k(y) = \mathcal{I}_{\{0\}}(y)$. Then, we have the following formulation for problem (P):

$$
(\hat{y}^{(1)}, \cdots, \hat{y}^{(m)}) = \underset{\tilde{y}^{(i)}, i=1,\ldots,m}{\operatorname{argmin}} \sum_{i=1}^{m} f_i(\tilde{y}^{(i)}) + \sum_{i=1}^{m-1} h_i(\tilde{y}^{(i)} - \tilde{y}^{(i+1)}) \tag{P1}
$$

Note that the functions f_i and h_i we defined above satisfy the conditions of Lemma E.2. Thus, it only remains to find the bounds on the variables $\tilde{y}^{(i)}$. This is easy to achieve as all $\tilde{y}^{(i)}$ must satisfy $|\tilde{y}^{(i)}| \leq b|z_i| = b|\mathbf{w} \cdot \mathbf{x}^{(i)}|$ and we know that $\mathbf{x}^{(i)}$ are sub-exponential random variables. Therefore, following the same idea from the proof of Lemma F.2, we know that for $U = \frac{2W}{L} \log(m/(L\delta))$, it holds that with probability at least $1 - \delta$, $|\tilde{y}^{(i)}| \leq b|\mathbf{w} \cdot \mathbf{x}^{(i)}| \leq bU$ for all $i \in [m]$. Hence, applying Lemma E.2 to problem (P1), we get that it can be solved within ϵ-error with computation complexity $\tilde{O}(m^2 \log(bW/(L\epsilon)))$. □

Since the solution $\hat{v}^t$ is ϵ-close to $\hat{u}^t$, this approximated solution will only result in an ϵ-additive error in the sharpness result Proposition 3.1 and the gradient norm concentration Lemma 4.3. In more details, for the result of Proposition 3.1, we have

$$
\begin{aligned}
\left| (\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{v}^t) - \nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)) \cdot (\mathbf{w}^t - \mathbf{w}^*) \right| &= \left| \frac{1}{m} \sum_{i=1}^{m} (\hat{v}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - \hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}))(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)} \right| \\
&\leq \frac{\epsilon}{m} \sum_{i=1}^{m} \left| (\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)} \right| \leq 2\epsilon U,
\end{aligned}
$$

since $|\mathbf{w}^t \cdot \mathbf{x}^{(i)}| \leq U$ and $|\mathbf{w}^* \cdot \mathbf{x}^{(i)}| \leq U$ with probability at least $1 - \delta$. Therefore, choosing $\epsilon' = \epsilon/U$ we have that Proposition 3.1 holds for approximate activations $\hat{v}^t$ with an additional ϵ error.

Let us denote the unit ball by $\mathbb{B}$. For the gradient norm concentration lemma Lemma 4.3, note that at any iteration t, it always holds that

$$
\|\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{v}^t) - \nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 = \max_{\mathbf{v} \in \mathbb{B}} \frac{1}{m} \sum_{i=1}^{m} (\hat{v}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - \hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)})) \mathbf{x}^{(i)} \cdot \mathbf{v} \leq \max_{\mathbf{v} \in \mathbb{B}} \frac{\epsilon}{m} \sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}|.
$$

Since $\mathbf{x}$ is isotropic and $\mathbf{v} \in \mathbb{B}$, we have $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[|\mathbf{x} \cdot \mathbf{v}|] \leq \sqrt{\mathbf{E}[(\mathbf{x} \cdot \mathbf{v})^2]} \leq 1$. Now since $|\mathbf{x}^{(i)} \cdot \mathbf{v}|$ are independent $1/L$-sub-exponential random variables, applying Bernstein's inequality it holds that for any $\mathbf{v} \in \mathbb{B}$,

$$
\mathbf{Pr}\left[\left| \frac{1}{m} \sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}| - \underset{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}{\mathbf{E}} [|\mathbf{x} \cdot \mathbf{v}|] \right| \geq s \right] \leq 2 \exp\left(- c \min\left\{ \frac{m^2 s^2}{m/L^2}, \frac{ms}{1/L} \right\} \right) = 2 \exp(-cmL^2 s^2).
$$

Let $N(\mathbb{B}, \epsilon; \ell_2)$ be the ϵ-net of the unit ball $\mathbb{B}$. Note that the cover number of these $\mathbf{v} \in \mathbb{B}$ is of order $(1/\epsilon)^{O(d)}$, therefore, applying a union bound on $N(\mathbb{B}, \epsilon; \ell_2)$ and for all $t_0 JT = O(\log(1/\epsilon)/\sqrt{\epsilon})$ iterations, and setting $s = 1$, it holds

$$\mathbf{Pr}\left[\forall \mathbf{v} \in N(\mathbb{B}, \epsilon; \ell_2), \left|\frac{1}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}| - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{x} \cdot \mathbf{v}|]\right| \geq 1\right] \leq 2\exp(-cmL^2 + c'd\log(1/\epsilon)) \leq \delta,$$

where the last inequality comes from the fact that we have $m \gtrsim W^{9/2}b^{14}d\log(1/\delta)\log^4(d/\epsilon)/(L^4\mu^9\delta\epsilon^{3/2})$ as the batch size. Let $\mathbf{v}^* = \mathrm{argmax}_{\mathbf{v}\in\mathbb{B}} \sum_{i=1}^m |\mathbf{x}^{(i)} \cdot \mathbf{v}|$, there exists a $\mathbf{v}' \in N(\mathbb{B}, \epsilon; \ell_2)$ such that $\|\mathbf{v}' - \mathbf{v}^*\|_2 \leq \epsilon$ and hence,

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}^*| &\leq \frac{1}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot (\mathbf{v}^* - \mathbf{v}')| + \frac{1}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}'| \\
&= \frac{\epsilon}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \frac{\mathbf{v}^* - \mathbf{v}'}{\epsilon}| + \frac{1}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}'| \\
&\leq \frac{\epsilon}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}^*| + \frac{1}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}'|,
\end{aligned}$$

where the last inequality comes from the observation that as $(\mathbf{v}^* - \mathbf{v}')/\epsilon \leq \mathbb{B}$, it holds $\sum_{i=1}^m |\mathbf{x}^{(i)} \cdot ((\mathbf{v}^* - \mathbf{v}')/\epsilon)| \leq \sum_{i=1}^m |\mathbf{x}^{(i)} \cdot \mathbf{v}^*|$, by the definition of $\mathbf{v}^*$. Therefore, with probability at least $1 - \delta$ we have

$$\frac{1}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}^*| \leq \frac{1}{1-\epsilon}\frac{1}{m}\sum_{i=1}^{m} |\mathbf{x}^{(i)} \cdot \mathbf{v}'| \leq 2(1 + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{v} \cdot \mathbf{x}|]) \leq 4.$$

This implies that $\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{v}^t)\|_2 \leq \|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 + 4\epsilon$ for all iterations with probability at least $1 - \delta$. Therefore, Lemma 4.3 continues to hold for the approximately optimal activation $\hat{v}^t$ with only an additive ϵ error.

Thus, we have the inequations (46) and (48) in the proof of Theorem 4.2 remains valid for $\hat{v}^t$ with an additional ϵ error, and hence the results in Theorem 4.2 is unchanged.

F. Uniform Convergence of Activations

In this section, we provide some standard uniform convergence results showing that the empirical optimal activation concentrates nicely around the population activations. We first bound the L_2^2 distance between the sample-optimal and population-optimal activations under $\mathbf{w}^t$. To do so, we build on Lemma 8 in (Kakade et al., 2011). Note that Lemma 8 from (Kakade et al., 2011) only works for bounded 1-Lipschitz activations $u : \mathbb{R} \mapsto [0, 1]$, hence it is not directly applicable to our case. Fortunately, since $\mathcal{D}_\mathbf{x}$ has a sub-exponential tail (see Definition 1.2), we are able to bound the range of $u(\mathbf{w} \cdot \mathbf{x})$ for $u \in \mathcal{U}_{(a,b)}$ and $\mathbf{w} \in \mathbb{B}(W)$ with high probability. Concretely, we prove the following lemma. Note that in the lemma statement, $\hat{u}^{*t}$ is a random variable defined w.r.t. the (random) dataset S^*, and thus the probabilistic statement is for this random variable.

We make use of the following fact from (Kakade et al., 2011):

Fact F.1 (Lemma 8 (Kakade et al., 2011))**.** Let $\mathcal{V}$ be the set of non-deceasing 1-Lipschitz functions such that $v : \mathbb{R} \to [0, 1]$, $\forall v \in \mathcal{V}$. Given $S_m = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$, where $(\mathbf{x}^{(i)}, y^{(i)})$ are sampled i.i.d. from some distribution $\mathcal{D}'$, let

$$\hat{v}_\mathbf{w} \in \mathop{\mathrm{argmin}}_{v\in\mathcal{V}} \frac{1}{m}\sum_{i=1}^{m} (v(\mathbf{w} \cdot \mathbf{x}^{(i)}) - y^{(i)})^2.$$

Then, with probability at least $1 - \delta$ over the random dataset S_m, for any $\mathbf{w} \in \mathbb{B}(W)$ it holds uniformly that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}[(\hat{v}_\mathbf{w}(\mathbf{w} \cdot \mathbf{x}) - y)^2] - \inf_{v\in\mathcal{V}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}[(v(\mathbf{w} \cdot \mathbf{x}) - y)^2] = O\left(W\left(\frac{d\log(Wm/\delta)}{m}\right)^{2/3}\right).$$

The first lemma states that with sufficient many of samples, the idealized empirical activation $\hat{u}^{*t}$ defined as the optimal solution of (P*) is close to its population counterpart u^{*t}, the optimal solution of (EP*).

Lemma F.2 (Approximating Population-Optimal Noiseless Activation by Empirical)**.** *Let $\mathcal{D}_{\mathbf{x}}$ be (L, R)-well behaved and let $\mathbf{w}^t \in \mathbb{B}(W)$. Provided a dataset $S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}$, where $\mathbf{x}^{(i)}$ are i.i.d. samples from $\mathcal{D}_{\mathbf{x}}$ and $y^{*(i)} = u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$, let $\hat{u}^{*t}$ be the sample-optimal activation on S^* as defined in (P*). In addition, let u^{*t} be the corresponding population-optimal activation, following the definition in (EP*). Then, for any $\epsilon, \delta > 0$, if the size m of the dataset S^* is sufficiently large*

$$m \gtrsim d \log^4(d/(\epsilon\delta)) \left(\frac{b^2 W^3}{L^2 \epsilon}\right)^{3/2},$$

we have that with probability at least $1 - \delta$, for any $\mathbf{w}^t \in \mathbb{B}(W)$:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] + \epsilon\,,$$

and, furthermore,

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \le \epsilon.$$

Proof. Our goal is to show that with high probability, the empirical optimal activation $\hat{u}^{*t} \in \mathcal{U}_{(a,b)}$ and the population optimal activation $u^{*t} \in \mathcal{U}_{(a,b)}$ can be scaled to 1-Lipschitz functions mapping $\mathbb{R}$ to $[0, 1]$, then, Fact F.1 can be applied.

Since $\mathbf{x}$ possesses a sub-exponential tail, for any $\mathbf{w} \in \mathbb{B}(W)$ we have $\mathbf{Pr}[|\mathbf{w} \cdot \mathbf{x}| \ge \|\mathbf{w}\|_2 r] \le \frac{2}{L^2}\exp(-Lr)$. Therefore, with probability at least $1 - (\delta_1/m)^2$ it holds $|\mathbf{w} \cdot \mathbf{x}| \le \frac{2W}{L}\log(m/(L\delta_1))$. Since we have m samples, a union bound on these m samples yields that with probability at least $1 - \delta_1^2/m$ it holds $|\mathbf{w} \cdot \mathbf{x}^{(i)}| \le \frac{2W}{L}\log(m/(L\delta_1))$, for any given $\mathbf{w} \in \mathbb{B}(W)$. Let $r = \frac{2W}{L}\log(m/(L\delta_1))$. In the remaining of the proof, we assume that $\mathbf{w}^t \cdot \mathbf{x}^{(i)} \le r$ holds for every $\mathbf{x}^{(i)}$ in the dataset S^*, which happens with probability at least $1 - \delta_1^2/m \ge 1 - \delta_1$.

Let $\mathcal{V}$ be the set of non-decreasing 1-Lipschitz functions $v : \mathbb{R} \to [0, 1]$ such that $v(0) = 1/2$, and $v(z_1) - v(z_2) \ge (a/(2br))(z_1 - z_2)$ for all $z_1 \ge z_2 \ge 0$. We observe that restricted on the interval $|z| \le r$, $(\hat{u}^{*t}(z)/(2br) + 1/2)|_{|z|\le r}$ is 1-Lipschitz, non-decreasing and bounded in the interval $[0, 1]$. Thus, $(\hat{u}^{*t}(z)/(2br) + 1/2)|_{|z|\le r} = \hat{v}^{*t}(z)|_{|z|\le r}$, for some $\hat{v}^{*t} \in \mathcal{V}$. Furthermore, under the condition that $|\mathbf{w}^t \cdot \mathbf{x}^{(i)}| \le r$, since $(\hat{u}^{*t}(z)/(2br) + 1/2)|_{|z|\le r} = \hat{v}^{*t}(z)|_{|z|\le r}$, we observe that $v^{*t}(z)$ is the optimal activation in the function space $\mathcal{V}$, given the dataset S^* and parameter $\mathbf{w}^t$, i.e.,

$$\hat{v}^{*t} \in \mathop{\mathrm{argmin}}_{v\in\mathcal{V}} \frac{1}{m}\sum_{i=1}^{m}(v(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - (u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})/(2br) + 1/2))^2.$$

In other words, $\hat{u}^{*t}(z)/(2br) + 1/2$ is the optimal empirical activation in the function class $\mathcal{V}$ when restricted to the interval $|z| \le r$. Consider $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$, it holds that $\mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| \ge r] \le (\delta_1/m)^2$. Then, for any $\mathbf{w}^t \in \mathbb{B}(W)$, the expectation $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]$ can be decomposed into the following terms

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \le r\}] \\
&\quad + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| > r\}] \\
&\le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \le r\}] \\
&\quad + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2 + (u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| > r\}]\,.
\end{aligned}$$

Since both $\hat{u}^{*t}$ and u^* are (a, b)-unbounded functions such that $\hat{u}^{*t}(0) = u^*(0) = 0$, hence it holds $(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2 \le b^2W^2((\mathbf{w}^t/\|\mathbf{w}^t\|_2) \cdot \mathbf{x})^2$ and similarly, $(u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \le b^2W^2((\mathbf{w}^*/\|\mathbf{w}^*\|_2) \cdot \mathbf{x})^2$. Furthermore, since for any unit vector $\mathbf{a}$, the random variable $\mathbf{a} \cdot \mathbf{x}$ follows a $1/L$-sub-exponential distribution for $\mathcal{D}_{\mathbf{x}}$ is (L, R)-well behaved, thus, it holds that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\mathbf{a} \cdot \mathbf{x})^4] \le c/L^4$ for some absolute constant c. Therefore, after applying Cauchy-Schwarz to $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \ge r\}]$ it holds

$$\begin{aligned}
\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \ge r\}] &\le b^2W^2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[((\mathbf{w}^t/\|\mathbf{w}^t\|_2) \cdot \mathbf{x})^4]\,\mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| \ge r]} \\
&\le cb^2W^2\delta_1/(L^2m), \qquad (60)
\end{aligned}$$

and similarly, $\mathbf{E}[(u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \geq r\}] \leq cb^2W^2\delta_1/(L^2m)$. Thus, bringing back to the upper bound on $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]$ displayed above, we get

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \leq r\}] \\ + 2cb^2W^2\delta_1/(L^2m).\end{aligned}$$

We are now ready to apply Fact F.1 (note that $\mathcal{V}$ is a smaller function class compared to the class of 1-Lipschitz functions described Fact F.1, hence the results in Fact F.1 applies). Denote $A = \{\mathbf{x} : |\mathbf{w}^t \cdot \mathbf{x}| \leq r\}$. Let $y' = y^*/(2br) + 1/2$, $y^* = u^*(\mathbf{w}^* \cdot \mathbf{x})$. Since conditioning on the A, $\hat{u}^{*t}(z)/(2br) + 1/2$ is the optimal empirical activation, applying Fact F.1 we get with probability at least $1 - \delta_2$:

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x})/(2br) + 1/2 - (u^*(\mathbf{w}^* \cdot \mathbf{x})/(2br) + 1/2))^2|A] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{v}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - y')^2|A] \\ &\leq \inf_{v\in\mathcal{V}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(v(\mathbf{w}^t \cdot \mathbf{x}) - y')^2|A] + \tilde{O}(W(d\log(m/\delta_2)/m)^{2/3}).\end{aligned}$$

Let $\mathcal{V}|_{|z|\leq r}$ and $\mathcal{U}_{(a,b)}|_{|z|\leq r}$ be the functions from $\mathcal{V}$ and $\mathcal{U}_{(a,b)}$ restricted on the interval $|z| \leq r$, respectively. It's easy to see that by definition of $\mathcal{U}_{(a,b)}$ and $\mathcal{V}$, $(\mathcal{U}_{(a,b)}|_{|z|\leq r})/(2br) + 1/2 \subset \mathcal{V}|_{|z|\leq r}$. Therefore,

$$\begin{aligned}\inf_{v\in\mathcal{V}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(v(\mathbf{w}^t \cdot \mathbf{x}) - y')^2|A] &\leq \inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t \cdot \mathbf{x})/(2br) + 1/2 - y')^2|A] \\ &\leq \frac{1}{4b^2r^2} \inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t \cdot \mathbf{x}) - y^*)^2|A].\end{aligned}$$

Hence, with probability at least $1 - \delta_2$,

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{A\}] \\ &= 4b^2r^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{v}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - y')^2|A]\,\mathbf{Pr}[A] \\ &\leq 4b^2r^2 \inf_{v\in\mathcal{V}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(v(\mathbf{w}^t \cdot \mathbf{x}) - y')^2|A]\,\mathbf{Pr}[A] + \tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3})\,\mathbf{Pr}[A] \\ &\leq \inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{A\}] + \tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3}) \\ &\leq \inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] + \tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3})\ .\end{aligned}$$

Setting $\delta_1 = \delta_2 = \delta/2$ and plugging everything back to (60), we finally get that with probability at least $1 - \delta$,

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \\ &\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \leq r\}] + 2cb^2W^2\delta/(L^2m) \\ &\leq \inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] + O\left(\frac{b^2W^3}{L^2}\log^2\left(\frac{m}{L\delta}\right)\left(\frac{d\log(m/\delta)}{m}\right)^{2/3}\right).\end{aligned}$$

To complete the first part of the claim, it remains to choose m as the following value

$$m = \Theta\left(d\log^4(d/(\epsilon\delta))\left(\frac{b^2W^3}{L^2\epsilon}\right)^{3/2}\right).$$

For the second part of the claim, note that $\mathcal{U}_{(a,b)}$ is a closed convex set of functions, and that the infimum $\inf_{u\in\mathcal{U}} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]$ is attained by $u^{*t}(z)$. As we have shown that with the sample size m specified above, with probability at least $1 - \delta$, it holds

$$\begin{aligned}\epsilon &\geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 - (u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) + u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - 2u^*(\mathbf{w}^* \cdot \mathbf{x}))] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))].\end{aligned}$$

Since $\hat{u}^{*t}(z) \in \mathcal{U}_{(a,b)}$, applying the second part of Claim C.6 with $v' = \hat{u}^{*t}$ we get

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - \hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(u^*(\mathbf{w}^*\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))] \geq 0.$$

Thus, we have:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] \leq \epsilon.$$

This completes the proof of Lemma F.2. □

To prove a similar uniform convergence result for the attainable activations $\hat{u}^t$, we make use of the following fact from prior literature, which shows that we can without loss of generality take the noisy labels to be bounded by $M = O(\frac{bW}{L}\log(bW/\epsilon))$, due to $\mathcal{D}_\mathbf{x}$ being (L,R)-well behaved.

Fact F.3 (Lemma D.8 (Wang et al., 2023))**.** Let $y' = \mathrm{sign}(y)\min(|y|, M)$ for $M = \frac{bW}{L}\log(\frac{16b^4W^4}{\epsilon^2})$. Then:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^*(\mathbf{w}^*\cdot\mathbf{x}) - y')^2] = \mathrm{OPT} + \epsilon.$$

In other words, we can assume $|y| \leq M$ without loss of generality by truncating labels that are larger than M. Under this assumption, as stated in Lemma F.4 below, we bound the L_2^2 distance between $\hat{u}^t$ and u^t using similar arguments as in Lemma F.2.

Lemma F.4 (Approximating Population-Optimal Activation by Empirical)**.** *Let $\mathbf{w}^t \in \mathbb{B}(W)$. Given a distribution $\mathcal{D}$ whose marginal $\mathcal{D}_\mathbf{x}$ is (L,R)-well behaved, let $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$, where $(\mathbf{x}^{(i)}, y^{(i)})$ for $i \in [m]$ are i.i.d. samples from $\mathcal{D}$. Let $\hat{u}^t$ be a sample-optimal activation for the dataset S and parameter vector $\mathbf{w}^t$, as defined in (P). In addition, let u^t be the corresponding population-optimal activation, as defined in (EP). Then, for any $\epsilon, \delta > 0$, choosing a sufficiently large*

$$m \gtrsim d\log^4(d/(\epsilon\delta))\left(\frac{b^2W^3}{L^2\epsilon}\right)^{3/2},$$

we have that for any $\mathbf{w}^t \in \mathbb{B}(W)$, with probability at least $1-\delta$ over the dataset S:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - y)^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - y)^2] + \epsilon\,,$$

and, furthermore,

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2] \leq \epsilon.$$

Proof. As in the proof of Lemma F.2, we choose $r = \frac{2cW}{L}\log(m/(L\delta_1))$ so that $|\mathbf{w}^t\cdot\mathbf{x}^{(i)}| \leq r$ for all $\mathbf{x}^{(i)}$'s from the dataset with probability at least $1 - \delta_1^2/m \geq 1 - \delta_1$. We now condition on the event that $|\mathbf{w}^t\cdot\mathbf{x}^{(i)}| \leq r$ for all $i = 1,\ldots,m$. Let $\mathcal{V}$ be the set of non-decreasing 1-Lipschitz functions such that $\forall v \in \mathcal{V}$, $v(0) = 1/2$, and $v(z_1) - v(z_2) \geq (a/(2br))(z_1 - z_2)$ for all $z_1 \geq z_2 \geq 0$. Then, conditioned on this event, we similarly have that $(\hat{u}^t(z)/(2br) + 1/2)|_{|z|\leq r} = \hat{v}^t(z) \in \mathcal{V}$, and $\hat{v}^t(z)$ satisfies:

$$\hat{v}^t(z) \in \mathop{\mathrm{argmin}}_{v\in\mathcal{V}} \frac{1}{m}\sum_{i=1}^m (v(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - y^{(i)})^2.$$

Again, studying the L_2^2 distance between $\hat{u}^t(z)$ and $u^t(z)$, we have:

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - y)^2] = &\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - y)^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}| \leq r\}] \\ &+ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - y)^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}| > r\}].\end{aligned}$$

The probability of $|\mathbf{w}^t \cdot \mathbf{x}| > r$ is small due to the fact that $\mathcal{D}_\mathbf{x}$ possesses sub-exponential tail: $\mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| > r] \leq (\delta_1/m)^2$. Now note that $|y| \leq M$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t/\|\mathbf{w}^t\|_2) \cdot \mathbf{x})^4] \leq c/L^4$ by the sub-exponential property of $\mathcal{D}_\mathbf{x}$, we thus have:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| > r\}] \\
&\leq 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[((\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}))^2 + y^2) \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| > r\}] \\
&\leq 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[b^2 W^2 ((\mathbf{w}^t/\|\mathbf{w}^t\|_2) \cdot \mathbf{x})^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| > r\}] + 2M^2 \mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| > r] \\
&\leq 2b^2 W^2 \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{w}^t/\|\mathbf{w}^t\|_2) \cdot \mathbf{x})^4] \mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| > r]} + 2M^2 \mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| > r] \\
&\leq 2cb^2 W^2 \delta_1/(L^2 m) + 2M^2 (\delta_1/m)^2,
\end{aligned}$$

where in the second inequality we used the fact that $\hat{u}^t$ is b-Lipschitz and $\mathbf{w}^t \in \mathbb{B}(W)$, and in the third inequality we applied Cauchy-Schwarz. Since $M = \frac{bW}{L} \log(\frac{16b^4W^4}{\epsilon^2})$, we have $M^2(\delta_1/m) \lesssim cb^2W^2/L^2$ for $m \gtrsim \log(bW/\epsilon)$, thus, in the end, we get

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| > r\}] \leq 4c(bW/L)^2 \delta_1/m, \tag{61}$$

for some absolute constant c.

The rest remains the same as in the proof of Lemma F.2. Let $A = \{\mathbf{x} : |\mathbf{w}^t \cdot \mathbf{x}| \leq r\}$. Let $y' = y/(2br) + 1/2$. As $\hat{v}^t(z) = \hat{u}^t(z)/(2br) + 1/2$ is the optimal empirical activation in $\mathcal{V}$ given $\mathbf{w}^t$ (conditioned on A), applying Fact F.1 we have with probability at least $1 - \delta$:

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[((\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x})/(2br) + 1/2) - y')^2 | A] &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{v}^t(\mathbf{w}^t \cdot \mathbf{x}) - y')^2 | A] \\
&\leq \inf_{v\in\mathcal{V}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(v(\mathbf{w}^t \cdot \mathbf{x}) - y')^2 | A] + \tilde{O}(W(d\log(m/\delta_2)/m)^{2/3}).
\end{aligned}$$

Since $\mathcal{U}_{(a,b)}|_{|z|\leq r}/(2br) + 1/2 \subset \mathcal{V}|_{|z|\leq r}$, we further have

$$\begin{aligned}
\inf_{v\in\mathcal{V}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(v(\mathbf{w}^t \cdot \mathbf{x}) - y')^2 | A] &\leq \inf_{u\in\mathcal{U}_{(a,b)}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}^t \cdot \mathbf{x})/(2br) + 1/2 - y')^2 | A] \\
&\leq \frac{1}{4b^2r^2} \inf_{u\in\mathcal{U}_{(a,b)}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}^t \cdot \mathbf{x}) - y)^2 | A].
\end{aligned}$$

Therefore, $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2 \mathbb{1}\{A\}]$ can be bounded from above by

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2 \mathbb{1}\{A\}] \\
&= 4b^2r^2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{v}^t(\mathbf{w}^t \cdot \mathbf{x}) - y')^2 | A] \mathbf{Pr}[A] \\
&\leq 4b^2r^2 \inf_{v\in\mathcal{V}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(v(\mathbf{w}^t \cdot \mathbf{x}) - y')^2 | A] \mathbf{Pr}[A] + \tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3}) \\
&\leq \inf_{u\in\mathcal{U}_{(a,b)}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}^t \cdot \mathbf{x}) - y)^2 \mathbb{1}\{A\}] + \tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3}) \\
&\leq \inf_{u\in\mathcal{U}_{(a,b)}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] + \tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3}).
\end{aligned}$$

Thus, combining with (61), we get with probability at least $1 - \delta_1 - \delta_2$,

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] + \tilde{O}\left(Wb^2r^2\left(\frac{d\log(m/\delta_2)}{m}\right)^{2/3}\right) + \left(\frac{bW}{L}\right)^2 \frac{\delta_1}{m}.$$

Choosing the size of the sample set to be:

$$m = \Theta\left(d\log^4(d/(\epsilon\delta))\left(\frac{b^2W^3}{L^2\epsilon}\right)^{3/2}\right),$$

and recall that $r = \frac{2cW}{L}\log(m/(L\delta_1))$, we finally have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2] \le \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2]+\epsilon,$$

with probability at least $1-\delta$, after choosing $\delta_1=\delta_2=\delta/2$.

To prove the final claim of the lemma, we follow the same routine as in Lemma F.2. Since we have just shown that with probability at least $1-\delta$, it holds

$$\begin{aligned}\epsilon &\ge \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2-(u^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2]\\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2]+2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))(u^t(\mathbf{w}^t\cdot\mathbf{x})-y)],\end{aligned}$$

applying the first statement in Claim C.6 finishes the proof. □